FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Online Advertising: Forecasting and Synthesising Web Activity Based On Historical Data

Pedro Manuel Santos Borges

Mestrado Integrado em Engenharia Informática e Computação

Supervisor: João Mendes Moreira (PhD) - FEUP

Co-Supervisor: Hugo Sereno Ferreira (PhD) - FEUP

Company Supervisor: João Azevedo - ShiftForward

July 28, 2014

Online Advertising: Forecasting and Synthesising Web Activity Based On Historical Data

Pedro Manuel Santos Borges

Mestrado Integrado em Engenharia Informática e Computação

Approved in oral examination by the committee:

Chair: Doctor Ana Paula Cunha da Rocha
External Examiner: Doctor Nuno Filipe Fonseca Vasconcelos Escudeiro
Supervisor: Doctor João Pedro Carvalho Leal Mendes Moreira

July 28, 2014

Abstract

The online advertisement industry handles a large quantity of money and users everyday. This industry is always trying to get more efficient, for example, by enhancing the targeting of online advertising campaigns.

This pursuit of efficiency on the world of online advertising turned simpler methods of prediction unable to report an accurate number of impressions, used to calculate the value of a publisher's inventory. The introduction of concepts like frequency capping made that very clear.

There is now the need not only to predict the number of visits, but also to predict when this visits will happen, what the user did before going to that website and who he is.

In this document that concept will be approached using Data Mining techniques, such as clustering and time series analyses, in order to generate a future ad request log using only past data.

This generated results will be in the same format as the input dataset, to be used on simulators capable of calculate important metrics, for publishers and advertisers, for a set of campaigns.

Resumo

O mercado da publicidade online envolve diariamente muito dinheiro e muitos utilizadores. Este mercado que esta constantemente à procura de formas de se tornar mais eficiente, por exemplo, melhorando a selecção do público alvo das suas campanhas publicitárias.

Esta busca pela eficiencia no mercado da publicidade online tornou metodos de previsão mais simples incapazes de calcular correctamente o número de impressões, utilizadas para calcular o valor do inventario de um *publisher*. A introdução de conceitos como o *frequency capping* torna isso muito evidente.

Há actualmente a necessidade de não só prever o número de visitas, como também quando vão ocorrer essas visitas, o que o utilizador fez antes de lá chegar e quem é o utilizador em questão.

Neste trabalho esse conceito vai ser abordado recorrendo a técnicas de *Data Mining*, como o *clustering* e analise de series temporais, de forma a conseguir gerar um registo futuro de pedidos de publicidade utilizando apenas dados passados.

Os registos gerados estarão prontos a serem posteriormente utilizados, em simuladores capazes de calcular os resultados, para um universo de campanhas.

Acknowledgements

I want to direct my gratitude to my supervisor João Mendes Moreira for all the time we spend trying to mine the knowledge about data mining on my raw rock mind, without him none of this would be possible. I want to direct the same gratitude to the company supervisor João Azevedo, for all the insights about the online advertising market ecosystem. To my co-supervisor a special thanks to the inception of failure that you implanted on my brain which help me get to this point.

To all my friends on FEUP with whom I spend allmost every waking hour of my life for the past 5 years. A giant thanks to all of you, specially, Ana Ferreira, Bruno Maia, Daniel Freitas, Marco Amador, Margarida Pereira, Maria Barreira, Pedro Oliveira, Renato Rodrigues, Renato Marinho, Tiago Mota and Tiago Rodrigues. All of you had a special role in my life on the past years. A special acknowledge for Diogo Teixeira and Clara Sacramento was to be made, thank you for all the great times we spent at NIAEFEUP, without you I would not be able to complete this report.

To my parents António and Josefina, thank you for everything that you did so I could become the person I am today. I hope I could live to fulfill all the expectations you have for me.

Maria Soutelo, thank you for every hour you had to spend without me, thank you for every hour you had been kept wake so I could work, thank for all the help! Without you I would not be writing this right now. Thank you for making me the happiest person on the face of this planet.

Pedro Borges

"No! Marty! We've already agreed that having information about the future can be extremely dangerous. Even if your intentions are good, it can backfire drastically!"

Dr. Emmett Brown

Contents

List of Figures

List of Tables

List of Algorithms

Abbreviations

ad	advertisement
adtech	Advertising Technology
AR	Autoregressive
CPA	Cost-per-Action
CPC	Cost-per-Click
CPL	Cost-per-Lead
CPO	Cost-per-Order
CPM	Cost-per-Mile
DSP	Demand Side Platform
MA	Moving Average
MLP	Multilayer Perceptron
MR	Multiple Regression
RTB	Real Time Bidding
SSP	Supply Side Platform
SVM	Support Vector Machines

Chapter 1

Introduction

1.1 Context and Framing

The online marketing is a growing multibillion-dollar industry [Pri13a] which is expected to continue its fast growth[Pri13b].

This industry is always trying to become more efficient by getting more profit from assets it already owns. Web users are the major assets of this industry, which makes money by exploiting the user behavior and characteristics, in order to target them with the perfect campaign. Each campaign has its own target parameters, which limit the target user universe. Online marketing industry core business is centered in web users and this industry has recorded almost every footprint each user makes on the web. Future footprints of the web users allows the measurement of the behavior of an upcoming campaign and, with this data, it is possible to make the inventory more profitable. Therefore, using future user data allows the adtech industry to be able to fine tune its campaigns. Campaigns are composed by a series of ads that share the same main idea they want to transmit. The campaigns have a targeting typically defined as a set of parameter definitions and rules. To be able to run the campaigns in a simulator, their targeting can be defined as queries over the ad requests' data. The utilization of a simulation allows to get the results fast, test concurrent campaigns and test multiple scenarios. The utilization of a simulation is the main reason behind why is so important to be able to generate future ad requests' data.

The most common platforms that will benefit from this data are Custom-Built Ad Servers and Exchanges, Sell-Side Platforms (SSPs) and Demand-Side Platforms (DSPs). These platforms are further explored in section 2.1.

1.2 Project

The online advertising market is huge and its size has been increasing in money, campaigns and users. Both platforms that sell and buy space for ad placement want to understand

what is their value and more importantly, what will be their value in the future. In both cases this value is mostly constrained by campaigns and the users they want to target.

Our goal is to forecast the availability of the users in the future so we can simulate the value of future campaigns over them.

Since we do not know which characteristics the future campaigns will have, every detail available on the impressions needs to be forecasted, in order to obtain the correct values when the queries are executed over the generated data.

With this said, the main goal is to be able to generate a dataset able to be used on a campaign simulator, so we must be able to predict the values with the maximum detail possible.

This simulator needs to have available every detail possible about every impression in order to identify which impressions are compatible with each campaign. The result would be expressed by number of impressions per campaign and users target by every campaign, over time.

The approach should also only need an impression's date and user id, with all other variables being optional. This constraint is imposed by the multiple sources of the dataset used, since each one of these sources could store different details and different types of parameters about each impression, so we cannot rely on the availability of such parameters.

The approach should be able to generate data for any source with any parameters, based only on historical information.

To conclude, the approach's main goals are to:

- Correctly predict volumes of activity on an ad network for a given time in the future based only in past data;
- Fill the volumes with impressions, with the maximum detail possible, to be able to use the obtained result on a simulator.

1.3 Motivation and Goals

In the last few years, the online marketing has been getting more complex. In such a way that today campaigns have a very well defined target, with sets of rules and limitations. This poses a big problem to simpler prediction models that normally don't predict all the parameters of the ad request, this way limiting the parameters where queries can be done.

Nowadays, some online ads can only be imprinted if a set of very specific requirements has been fulfilled, for example, the users had to visit an e-commerce site in the last 24 hours. This brings causality into the equation, creating a new paradigm that makes the more traditional methods of prediction ineffective. To solve this problem and to be able to get fast responses to complex queries of concurrent campaigns, simulate the algorithms executed by ad servers of the client and ultimately parallelize the computation of the

results for the queries, the complete future data has to be predicted. This generated data can be used in simulations and the online campaigns can run on top of the future population.

The objective of this thesis is to develop a library capable of generating future ad request logs using past data from the same network. This library will have as one of its main goals the prediction of all the parameters that characterize an ad request with the purpose of being able to query over any parameter, in other words, the generated dataset (ad requests log) must have the same attributes as the original.

The prediction of this kind of future data is rather complex since it is necessary to find out which users will appear in the future and which websites they will visit and when will they do it.

1.4 Report Structure

Besides this first introductory chapter, this report is divided into four additional chapters. In Chapter 2 it is explained some basic knowledge about online marketing. In addiction, there is a small discussion on which of those methods are more adequate to help solve this problem. Chapter 3 describes the proposed approach and also explains how the experimental setup was configured and designed. Chapter 4 focus on the results and their analyses, it is on this chapter where the results of all the phases of the approach are evaluated. Chapter 5 sums up the report, giving a better context of all the review done in the final project, it also includes some suggestions for future work that can be done to improve the proposed solution.

Chapter 2

State of the Art

This chapter starts with an overview of online marketing in the last few years, followed by a review of the data mining algorithms that can be used to solve the same kind of problems as this thesis.

2.1 Online Advertising Overview

Before entering in details about the state of the art of the technologies that can be used to solve the presented problem, it is better to explain some basic concepts about the world of online advertising.

All advertising has the main purpose of getting a message to the people that will impact or influence them in some way, therefore the same goal is applied to online advertising. One of the metrics of advertising are impressions, which correspond to the number of times a user sees the message (the ad)[kOAa]. **Ads** can present itself in various sizes[kOAc], forms and locations [kOAb], and these characteristics are chosen both by the advertiser and the publisher to better serve their purpose. **Campaigns** are composed by two big parts, which are the ads that compose it and the target population that they pretend to reach, including the rules of this targeting. For example, **frequency capping** to limit the number of times the same advertising is shown to the user [kOAa], avoiding, in this way, showing the same ad multiple times in a row to the same user, that can lead to a bad response from his part[BFGN14].

Nowadays, the main pricing models of online advertising are:

- **Cost-per-Mile (CPM)** where the advertiser pays per impression. The main problem of this model is the advertiser must pay to the publisher even if the ad doesn't lead to any profit.

- **Cost-per-Click (CPC)** where the advertiser pays per click to the publisher. This model is more expensive per unit[Per], but on overall can be more profitable[Per] if the audience of the websites where the ad is imprinted is more interested in that kind of product/service[And04].

- **Cost-per-lead (CPL)** where the advertiser pays for a lead. If this model is being used the advertiser doesn't pay per number of impressions nor per clicks. Instead, pays only if he gets valid information about the user, like the information of a sign up form for a community.

- **Cost-per-Action (CPA)** or **Cost-per-Order (CPO)** where the advertiser is charged per buy or action. This model is similar to CPL but has in mind an instantaneous return of the investment.

Traditionally, publishers sell their space to advertisers in bulk (**Ad networks**) this method has its *ups* and *downs*. The obvious *up* is that sometimes the advertiser gets premium spots at low prices. On the other hand, one of the biggest drawbacks is that when the advertiser buys the impressions as a closed package, sometimes impressions are not maximized in terms of profit. Other problem of traditional methods that, although the *CPA* and *CPL* pricing methods minimize the risk for the advertiser, the responsibility of optimizing conversion rate[1] is still on the ad network hands[YWZ13].

In the past few years, a new model called **Real Time Bidding (RTB)** has been gaining terrain [Adf]. *RTB* is a market where publishers offer his advertisement space and advertisers bid over it in real time. This allow publishers to get the best value for their space and advertisers get the best placement for their advertisement.

There are three main players in the world of *RTB*:

- The **Demand Side Platform (DSP)** is a tool used by the advertisers to act on their behalf on the *RTB*. *DSPs* allows them to set their campaigns' parameters and to monitor the performance of the campaign. This way the advertisers try to get the best performance of their campaigns because *DSPs* use algorithms driven by performance data[Ger12].

- The **Publisher** provides the inventory, that is comprised by accesses made by users. In some cases, the publisher uses **Supply Side Platforms**. *SSPs* help the publisher to better manage his inventory, and even let him set a reserve price for their inventory[YWZ13].

- The **Ad Exchange** looks a little like a stock exchange, but in reality is a software platform that mediates the exchange. This exchange takes place in a few milliseconds while the page loads.

[1] See, e.g., http://www.marketingterms.com/dictionary/conversion_rate/

RTB allows some features of paid search advertising everywhere [Ger12], because it allows the advertiser to better select the inventory[2] where he wants their campaigns to run on. The flexibility that *RTB* gives to all the intervinients of this exchange is what demands the necessity of predicting the future inventory, to better access its value.

2.2 Data Mining

"Data mining is about solving problems by analyzing data already present in databases."[WF05, p. 5] . Furthermore, consists in a vast number of techniques used to find interesting patterns in large datasets and translate that huge quantity of raw data in information and/or knowledge.

Data mining uses techniques from various fields, mostly from mathematics and computer science, such as artificial intelligence, machine learning and statistics. Data mining is sometimes referred as the natural evolution of information technologies [HK06a, p. 1]

There are lots of methods of data mining, which can be separated in two groups: descriptive data mining and predictive data mining [FFPs+96]. The main focus of the first group is to find the underlying structure of a given dataset, which methods try to find relationships and connections between the values, without have the goal of predicting the future. On the other hand, predictive data mining goal is to predict explicit values from patterns found on the original data set. These methods are used to build models based on past events that can be used to predict future events. This division is not always sharp and in some cases an algorithm mixes the two methods (predictive & descriptive)[FFPs+96].

According to previous statements it is easy to notice that data mining doesn't apply only to one set of problems and can be used to solve many different types of problems. The most common family of problem types are:

- **Anomaly Detection** tries to discover abnormal data on the dataset. This can be useful for identifying suspicious activity on a bank account log for example.
- **Classification** aims to identify which of a given set of categories a new observation belongs to.
- **Clustering** aims to grouping similar data together, in a finite number of categories, without prior knowledge of the characteristics of each group or the data.
- **Dependency modeling** tries to find associations between variables. For example, trying to find out which clothes go well together.
- **Summarization** provides an overview of the dataset, sometimes including visual representation and/or report generation.

[2] this inventory is made of user accesses

- **Regression** tries to predict the value of a quantitative variable given a new observation.

Next, the families of algorithms that are more relevant to this problem will be described in more detail.

2.2.1 Classification Algorithms

2.2.1.1 Decision Trees

Decision trees algorithms use a decision tree as a predictive model where all internal nodes (non-leaf nodes) are a test for the value of an attribute that will ultimately lead to a leaf node with the class attribute value (see example in figure 2.1). In other words, the selection of the class value is only based on the attribute values of the entry.[HK06a]

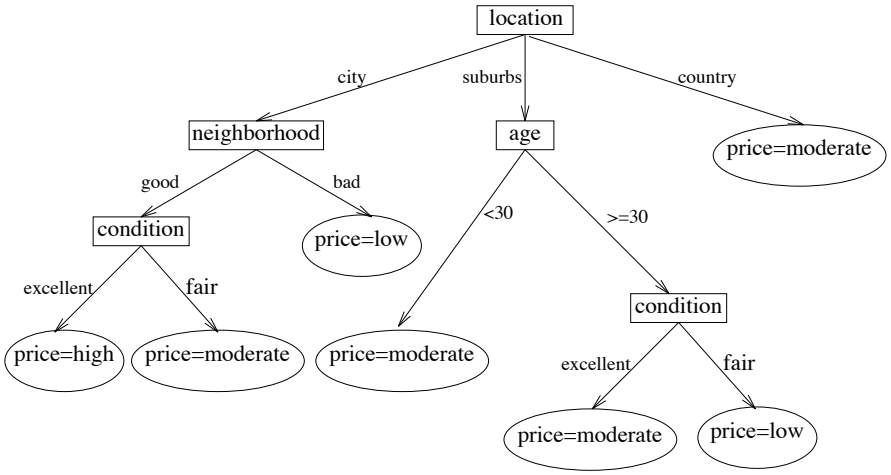

Figure 2.1: Example of a decision tree, Rectangles represent internal nodes and ovals represent leaf nodes (possible solution)[BA97]

Decision tree classifiers can be used in large datasets with high dimensional data and still be fast and easy to understand its result. One of its' main advantages due to its white box model of learning, is that at anytime it is possible to understand the reason behind each and every result. In addition to that, decision tree classifiers are very robust and require no prior knowledge of the domain or parameter settings.

On the other hand, the problem of learning an optimal decision tree is NP-Complete [HR76]. Therefore, in practical applications of decision tree learning algorithms, some heuristics need to be used, usually a greedy algorithm, which can lead to local optimal decisions being made for each node. The utilization of such algorithms cannot guarantee the global optimal solution to the problem. There are many algorithms that implement the decision tree principles, such as ID3, C5.0 and CART.

2.2.1.2 Random Forests

Random Forests [Bre01] are an ensemble learning method for classification and regression, that operates by generating a given number of decision trees from a randomly selected with replacement subset of the complete training dataset, where the subset distribution is the same across the forest. After that, at each node a randomly chosen subset of variables are used for the selection of the best split. This process continues until the trees are fully expanded. There is no pruning of the trees.

Random Forests are robust and fairly able to deal with unbalanced and missing data on the datasets. It is easy to set up with very little configuration parameters and it also gives good results even when the default parameters are used. The biggest limitation of this algorithm is not being able to predict beyond the range of the training data when used for regression, because randomly selected inputs give better results in classification than regression [Bre01].

2.2.1.3 Support Vector Machines

Support Vector Machine (SVM) is a supervised learning algorithm with great results in pattern recognition. [CV95] To achieve this results, SVMs rely on spatial division of classes. The division can be made without dimensional limit. In other words, the plane or hyperplane that separates the classes can have any number of dimensions. In figure 2.2 we can see an example of this multidimensional spatial division.

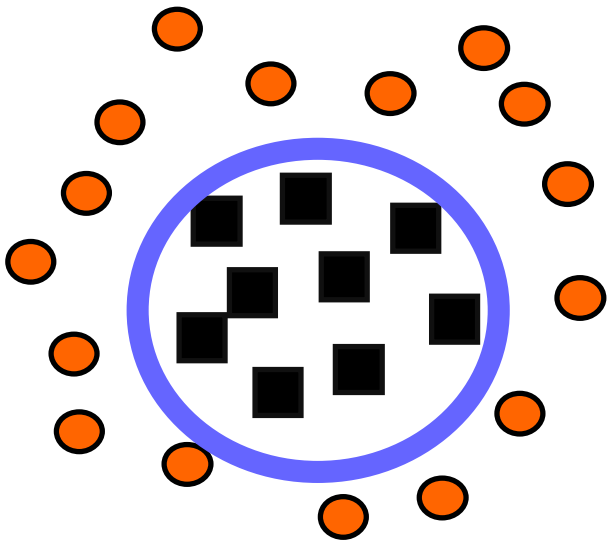

Figure 2.2: Example of a non linear separation (quadratic discriminant) [BC03]

However, the speed of training and testing is low, even for fastest SVMs. It is also directly limited to two-class tasks[CV95], requiring to reduce multi-class tasks to binary

ones. Although, some work has been done to try to avoid decomposing the problem in multiple binary class problems.[CS02]

In the past few years, this method is also being used to classify internet traffic with excellent results.[YLGX10]

2.2.1.4 KNN

The *k-nearest-neighbor* algorithm is among the most simple classification algorithms and it has been around since the 1950s [HK06a, p.348]. KNN is a non-parametric method which can be used in classification and regression problems and it is mostly used in the area of pattern recognition. This algorithm is a type of instance-based learning, or lazy learning, which means that every computation is deferred until classification. By deferring computation to the classification phase this algorithm is slow at classifying instances. To classify a new instance, the algorithm has to calculate the *Euclidean distance* between the new instance and every instance in the training set.

Since the introduction of this algorithm, some variations of it appeared to help solving the shortcomings of *kNN*.[BV10] The more notable are:

- **Clustered k nearest neighbor** [ZLX09] is an improved version of k-nn mixed with a clustering algorithm, which allow, for example, better performance in text classification.
- **k-d tree nearest neighbor (kdNN)** [Spr91] helps to improve, in some cases, the completion time of the classification in logarithmic time.
- **Orthogonal Search Tree Nearest Neighbor** [McN01] greatly improves efficiency of k-nn, especially for large datasets.

2.2.2 Clustering Algorithms

Clustering is the process of grouping similar entries/objects into different groups, in other words, it is the method of breaking a set into various subsets according to some metric. This groups/subsets are not known from the start and clustering is sometimes considered the most important unsupervised learning problem[Mad12].

Clustering methods can be divided into[HK06a]:

- **Partitioning Methods** start with an initial number of groups, and reallocates iteratively the elements on the groups to convergence[Mad12]. Some examples of partitioning methods are based on heuristics like *k-mean algorithm* and *k-medoids algorithm*.
- **Hierarchical Methods** work by grouping data into a tree of clusters[HK06a]. This methods can be further divided into two groups: *agglomerative* (bottom-up) and

divisive (top-down)[Mad12]. At the beginning of an *agglomerative* algorithm, each object is a cluster and these clusters merge with each other, to form less but larger clusters. The opposite occurs for *divisive* algorithms. The end condition for both is a distance threshold.[HK06a]

- **Density-Based Methods** were developed to find clusters with odd forms, relying on the premise that the clusters are located in high density areas that are separated from each other by low density zones [HK06a]. Some examples of algorithms which implement this method are *DBSCAN* and *SSN*.

- **Grid-Based Methods** uses a multiresolution grid data structure. It divides the object space into a finite number of cells, that form a grid structure, on which all of the operations for clustering are performed. This approach has a fast processing time, which is typically independent from the number of data objects but, on other hand, dependent on the number of cells per dimension on the object space[HK06a]. Some examples of algorithms which implement the rules of this method are *STING*, *WaveCluster* and *CLIQUE*.

- **Model-Based Clustering Methods** tries to understand the mathematical rule behind the data, in other words, it is an ”attempt to optimize the fit between the given data and some mathematical model”[HK06a, p. 429]. This method assumes that the data was generated with some underlying probability. Some examples of this algorithms are *Expectation-Maximization*, *Conceptual Clustering* and *Neural Networks*.

2.2.3 Time Series Prediction

One of the main problems, is to accurately predict the volumes of the inventory. Since the data has temporal information one possible approach is to use time series prediction to calculate future values of time series calculated from the datasets.

There are various methods to solve this kind of problems, bellow some confusions of studies that have been done in this area are given.

[WK06] compared three forecasting techniques *Support Vector Machines* (SVMs), *Multiple Regression* (MR) and *Multilayer Perceptron* (MLP), on power production values on multiple power plants. The *MR* outperformed the other two methods on predicting those values.

[Sab07] compared *ARIMA* with logistic regressions algorithms to predict traffic on three Egyptian intercity roads. The average annual, monthly and weekly daily traffic volumes were calculated using both logistic regression and *ARIMA* algorithms. They concluded that *ARIMA* outperforms the logistics regression methods in forecasting this traffic volumes.

To address our problem the *ARIMA* approach will be explored. More details about this approach are available on the section bellow.

2.2.3.1 Arima

ARIMA also known as Box-Jenkins, is a modelling and forecasting approach. It combines three processes, the *Autoregressive* (AR), differencing to strip off the integration (I) of the time series and *Moving Average* (MA).

Each one of these processes handle the random disturbance in its own way.[Sab07] The *AR* part of this approach is the linear regression of the series against one or more prior values of the series. A time series is suceptible to capture noise shocks on a noisy environment and it may memorize these shocks for a while, the *MA* term is used to capture the outcomes of these shocks in the future.[LWT05] The combination of this to terms compose the *ARMA* model.

The *ARMA* model assumes that the data is stationary, which is, the statistical propreties of the data doesn't change overtime. However, this assumption doesn't hold against most of real time series.[BJR13] So, the *Integration* process has introduced in order to remove the impact of non-stationary data by differencing.

The three processes, AR (p), I (d) and MA (q) are combined and compose the ARIMA (p, q, d) model.

2.2.4 Data Mining Tools

Today there are several free tools on the internet which can help us to test and use data mining techniques. Some of the most used will be presented bellow.

2.2.4.1 Weka

Weka[3] is a popular, open source, suite of machine learning software written in Java. It was created at the University of Waikato, New Zealand in 1997. It has an easy to use and comprehensive GUI with access to an enormous deck of machine learning algorithms.

2.2.4.2 Apache Mahout

Apache Mahout[5] is an open source machine learning library to build scalable machine learning libraries. Its core algorithms are implemented on top of map/reduce paradigm, to help in the scalability of the solution. Since it is easier to get a cluster of server than an ultra high frequency CPU, and the market trend is to develop many and multi core solution, it is always the best option for processing large quantities of data scalable software.

[3] Available at http://www.cs.waikato.ac.nz/~ml/weka/
[4] http://www.siliconafrica.com/wp-content/themes/directorypress/thumbs//weka.png
[5] Available at http://mahout.apache.org

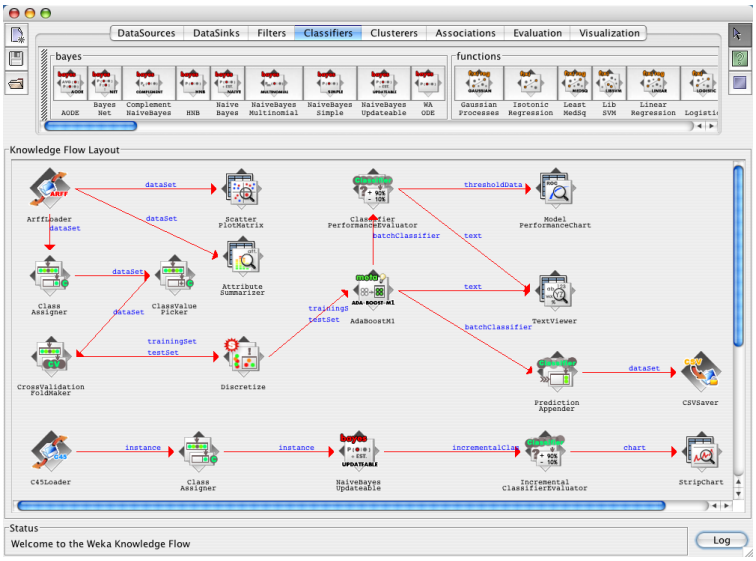

Figure 2.3: Screenshot of Weka [4]

2.2.4.3 RapidMiner

Rapid Miner[6] is a complete solution for data mining problems. It is available in a form of a standalone GUI application. Although it is a commercial product, there is also a free

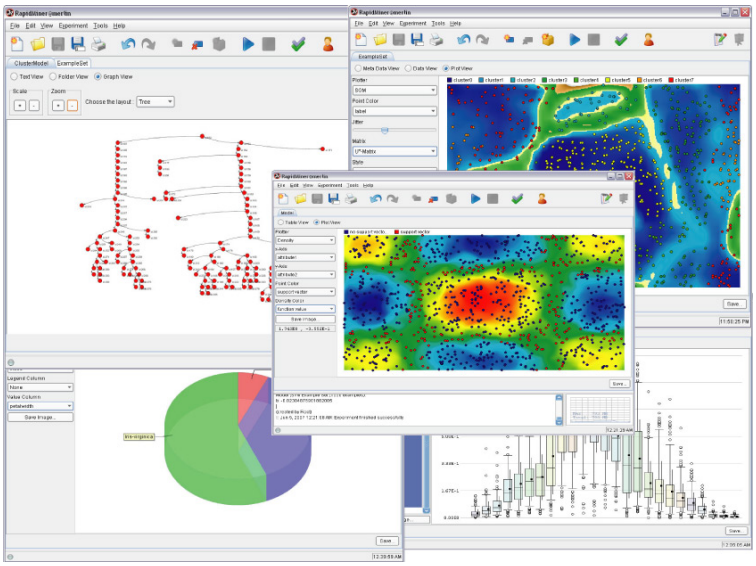

Figure 2.4: Screenshot of RapidMiner [7]

tier, and the core and earliest versions are open source. This application is one of the main players of this market and it is easily expandable through plug-ins available online.

[6] Available at http://www.rapidminer.com/
[7] http://mloss.org/media/screenshot_archive/rapidminer_collection.jpg

2.2.4.4 R Language

R[8] is a free programming language and software environment for graphics generation and statistical computing. Developed by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, in 1993 [Iha98], it is still in active development and greatly used by statistics and data miners.

R is an implementation of S, the statistics programming language, and it uses some characteristics inspired on Scheme. R is a GNU tool, so it is completely free. There are wrappers to almost every language which can be used to access R variables from other programming languages.

2.3 Model Evaluation Procedures and Measures

To verify if a created model is valid it is necessary an additional step to validate the data. This step can be divided into evaluation procedures, which normally consists of dividing the original dataset into training and testing subsets, and evaluation measures in order to assess the quality of our model.

2.3.1 Model Evaluation Procedures

2.3.1.1 Cross-Validation

This model consists of dividing the dataset in parts [WF05] . Some of these parts will be used to train the model and other parts will only be used in the validation part, so that we can assess if the model is able to generalize the result or not.

The training set is the only part of the dataset that the training algorithm uses to generate the model, the validation set will be used to calculate the error when comparing the real result with the result given by the generated model.

Sliding Window This procedure selects its training data and validation data maintaining the chronological order of the data set[BBR+07].

This procedure can only be used in time series validation since it depends on time points being present in the dataset. It uses a certain window of examples for training, and the examples that follow (in terms of time) as testing examples. The window is moved across the example set and all performance measurements are averaged afterwards.

[8] Available at http://www.r-project.org

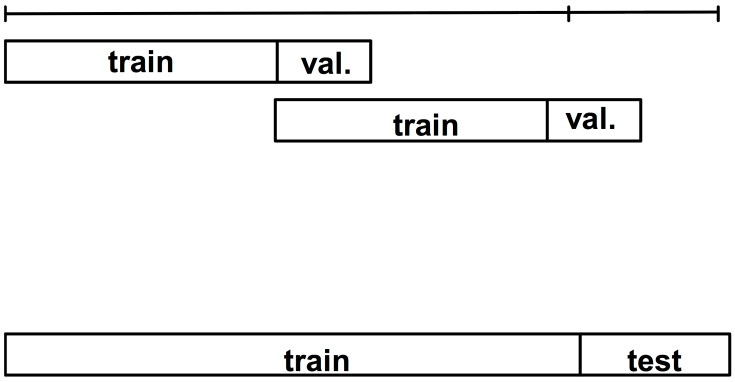

Figure 2.5: Sliding window validation

In figure 2.5, it possible to see that the validation data happens after the training data in a temporal point of view.

2.3.2 Model Evaluation Methods

2.3.2.1 Precision

Precision is a metric to identify the fraction of positives that are correctly classified as so. In other words, using this measure, we can identify how relevant the results given by the algorithm are (the more higher the value, the more relevant the results are). The equation to calculate precision is presented next:

$$Precision = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

2.3.2.2 Recall

This measure represents the part of actual positive results in the dataset that were identified as such. In other words, recall is the fraction of positive objects that were correctly identified by the algorithm from all the positive objects in the dataset. This metric is calculated using the following expression:

$$Recall = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

2.3.2.3 Accuracy

Accuracy represents the percentage of results that are actually correct. This measure can be calculated as follows:

$$Accuracy = \frac{\text{true positives} + \text{true negatives}}{\text{true positives} + \text{false positives} + \text{true negatives} + \text{false negatives}}$$

2.3.2.4 F-Measure

F-Measure or F-Score or F_1 measures the test accuracy relying in both recall and precision. The best value of this function is 1 and the worst is 0. F-Measure is the harmonic mean of the precision and recall and can be defined as it follows:

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

The reason behind the utilization of the harmonic mean instead of the arithmetic mean is given due to a more intuitive result [Sas07].

2.3.3 Time Series Forecast Accuracy Evaluation

There are various methods of accuracy evaluation for time series. In [HK06b] a large scale comparison of various methods, presented in Table 2.1, are described along this section. They are divided in five different groups: scale-dependent, based on percentage errors, based on relative errors, relative measures and lastly scaled errors. This section will help us to select the best method based on the presented state of the art.

Exhibit 9
Ratings of the error measures

Error measure	Reliability	Construct validity	Outlier protection	Sensitivity	Relationship to decisions
RMSE	Poor	Fair	Poor	Good	Good
Percent Better	Good	Fair	Good	Poor	Poor
MAPE	Fair	Good	Poor	Good	Fair
MdAPE	Fair	Good	Good	Poor	Fair
GMRAE	Fair	Good	Fair	Good	Poor
MdRAE	Fair	Good	Good	Poor	Poor

Table 2.1: Guidelines for selecting error measures [AC92]

2.3.3.1 Scale-dependent measures

These accuracy measures are commonly used whose scale depends on the scale of the data. So these methods shall only be used when comparing different methods applied to the same data set, and shall never be used when comparing across different data sets whose have different scales.

Forecast error can be defined as $e_t = Y_t - F_t$, let Yt denote real observation at time t and Ft being the forecast value.

- **Mean Squared Error** (MSE) = mean(e_t^2)

- **Root Mean Square Error** (RMSE) $= \sqrt{MSE}$
- **Mean Absolute Error** (MAE) $=$ mean$(|e_t|)$
- **Median Absolute Error** (MdAE) $=$ median$(|e_t|)$

Some authors [Arm01] recommend the usage of *MAE* and *MdAE* instead of *MSE* and *RMSE* since the last are more susceptible to outliers than the first ones.

2.3.3.2 Measures based on percentage errors

The percentage error have the advantage of being scale independent thus can be used to compare forecast performance across different datasets.

The percentage error is given by $p_t = 100 * \frac{e_t}{Y_t}$.

The most common measures are:

- **Mean Absolute Percentage Error** (MAPE) $=$ mean$(|p_t|)$
- **Median Absolute Percentage Error** (MdAPE) $=$ median$(|p_t|)$
- **Root Mean Square Percentage Error** (RMSPE) $= \sqrt{mean(p_t^2)}$
- **Root Median Square Percentage Error** (RMdSPE) $= \sqrt{median(p_t^2)}$

The major drawback of these measures is for $Y_t = 0$ the value is infinite or undefined for any t and have a very skewed distribution when compared when any value of Y_t is close to 0. The *MAPE* and *MdAPE* also have the disadvantage of put a heavier penalty on positive errors than on negative errors. This observation lead to the use of another percentage error based measures. The "symmetric" measures [Mak93]:

- **Symmetric Mean Absolute Percentage Error** (sMAPE) $=$ mean$(200 * \frac{|Y_t - F_t|}{Y_t + F_t})$
- **Symmetric Median Absolute Percentage Error** (sMdAPE) $=$ median$(200 * \frac{|Y_t - F_t|}{Y_t + F_t})$

In case of the time series used have negative values, there are the risk of divide by zero. Furthermore these measures also are reported as not as "symmetric" as the name suggests. Mostly because the penalty is heavier when the forecasts are low when compared with when the forecasts are high. [GL99]

2.3.3.3 Measures based on relative errors

Another approach for scaling is to divide each error by the error obtained by another standard method of forecasting. So relative error is defined by $r_t = \frac{e_t}{e_t^*}$, where e_t^* is the forecast error obtained from the benchmark method. Usually this baseline method is the random walk where F_t is equal to the value of the last observation. So the measures can be defined:

- **Mean Relative Absolute Error** (MRAE) = mean($|r_t|$)
- **Median Relative Absolute Error** (MdRAE) = median($|r_t|$)
- **Geometric Mean Relative Absolute Error** (GMRAE) = gmean($|r_t|$)

2.3.3.4 Relative measures

As a substitute of the usage of relative error is the usage of relative measures. One example, is calculate the *MAE* of the baseline (MAE_b) and use the result to calculate the relative *MAE* defined by:

$$RelMAE = \frac{MAE}{MAE_b}$$

Similar approaches can be applied to other measures described before. The main advantages of these approaches is their interpretability, since if $RelMAE<1$ then the proposed method is better than the baseline and when the $RelMAE>1$ then the proposed method is worst than the baseline. A related approach is to use the percentage of forecasts which a given model is better than the baseline method.[AC92] Can be defined as:

$$PB_m = \frac{\sum_{t=1}^{n} j_t}{n} * 100$$

$$\text{where } j_t = \begin{cases} 1 \text{ if } e_{t,m} < e_{t,b} \\ 0 \text{ otherwise} \end{cases}$$

This approach describes a method *m* with a *PB* score relative to the baseline.

2.3.3.5 Scaled errors

The author of the paper [HK06b] propose this approach to address the situations where more traditional scaled accuracy measures fail (as described before). So the scaled error is described as it follows:

$$q_t = \frac{e_t}{\frac{1}{n-1} * \sum_{i=2}^{n} |Y_i - Y_{i-1}|}$$

A scaled error is less than one if the method forecast is better than the average one-step naïve forecast computed in sample. Otherwise, if the error is greater than one the method is worst than the naïve approach. So the **Mean Absolute Scaled Error** (MASE) is defined simply by

$$MASE = mean(|qt|)$$

In a similar manner the results of *MASE* for a method, if less than one the method is better than the naïve approach and worse otherwise. Related methods such as *Root Mean Squared Scaled Error* (RMSSE) and *Median Absolute Scaled Error* (MdASE) can be defined in a similar manner. The only situation where this method falls short is when the historical data has always the same value since the denominator would be 0 and make the result of *MASE* infinite or undefined.

2.4 Web Usage Mining

The main area of this project on is the utilization of ad requests logs to predict future requests of the users. This is analogous to work which have been done in the area of user future requests prediction. Next, it will be presented a little overview of what has been studied in the area of web usage data mining in the past years, in order to identify possible approaches to solve the problem of this thesis.

In the area of pre-fetching web pages [NKM01], it has been proposed a new algorithm, WM_0, which takes into account the order between accesses and other specificities of the area. This algorithm gets good results for accuracy even when compared to other methods like *Prediction-by-Partial-Match* (PPM) and *Dependency Graph* (DG).

In another paper [GH03], three approaches to data mine from web logs are proposed. **Association Rules** (AR) is based on association rule learning which is a very popular data mining family of algorithms to find relationships between variables. The problem of finding web pages together in a web log is similar to that problem. **Frequent Sequences** (FS), is a technique that tries to find time ordered sequences of URLs that have been followed by past users. **Frequent Generalized Sequences** (FGS) involves the utilization of a generalized sequence, which is a sequence that allows wild cards in order to represent a user flow of navigation in a more flexible way. Some tests have been made by the author [GH03] to test the performance of this three methods using real web logs. According to the results given by FS, FS has better accuracy than AR and FGS.

In another paper, other authors propose a model that preserves the sequentiality of the clicks [FMK03]. The rules maintain the sequence of the click stream between the

antecedent and the consequent, to maintain sequentiality. This model also introduces the concept of temporality, which is reflected by the distance between the consequent and the antecedent by number of clicks to go from one to the other. This rule is very important because it allows not only to find which page is going to be accessed but also when will it be accessed. The proposed model, **Customizable Sequential Behavior Model**, can be adapted depending on the characteristics of the server, in order to capture the behavior of the users more accurately.

There is another approach that relies on the sequentially of the clicks [JL07]. This method uses the prefix set of web pages (pages that the user had already visited) to predict a postfix set (next pages that the user will visit). To select this pages, the algorithm uses a confidence threshold to select the pages to the postfix set from historical data.

An approach based on the **Longest Common Subsequence** (LCS) [JMMS] has also been used in web usage mining. In this case, the author proposes a two part architecture: an offline part where knowledge is extracted from the historical data and an online part where this knowledge is used to predict the next visits of the user. The prediction is done in the online part by appending the last request from the user to the history and using LCS. This architecture is implemented in **WebPum** [JMSM10] by the same authors and improves the accuracy by a little margin from the previous method, SUGGEST 3.0 [BS04], in the field of next page recommendations.

In other paper a method combining Markov model based sequential pattern mining with clustering is proposed [Ani10]. This combination gets about 12% more accuracy when compared with the traditional Markov model. The proposed models combine great accuracy from high order Markov Models with less space complexity from low order Markov models.

2.4.1 Web Usage Mining applied to Online Advertising

Data mining is also being used to predict the response of an user to an online ad[CML13]. Their main objective is to develop a framework that can predict the result of an user clicking in an advertising, based on the history that they have. The proposed framework uses Maximum Entropy[Nig99] because it is easy to implement, can be parallelized and scaled with respect to the number of features. It is easy to include model updates in this method. In this paper, they add a two-phase feature selection algorithm, to increase the automation and reduce the need of domain expertise (a generalized mutual information method to select the feature groups to be included and a feature hashing to have the ability of controlling the size of the models). To their experiments, the authors used logs with the same parameters as the datasets that will be used on this project.

In another approach to the same problem are introduced improvements in the context of traditional supervised learning based on FTRL-proximal online learning algorithm [MHS+13]. This paper also explores ways to save memory during the prediction, using filters to select features to be included in the model, such as *Poisson Inclusion* and *Bloom Filter Inclusion*, and they concluded that the method which allows better savings, without loosing much prediction accuracy, is *Bloom Filter Inclusion*. To solve the memory problem they also tested encoding values with fewer bits and other techniques.

In yet another paper about CTR(Click-Through-Rate) prediction [TOY+13], a two stage approach is used, the first stage being the construction of a ranking model with the clicked ad requests and then a sigmoid function converting the value of the ranking model into CTR. The method proposed achieves better results in terms of AUC, MSE and LogLoss when compared to: L2-loss linear SVM, logistic regression with only the clicked ad requests and logistic regression with all the ad requests.

In the area of forecasting ad impressions, a Bayes network was used to capture inter-dependencies between the query traffic features and the competitors in the auction[NMJ+13]. This method is used to forecast the number of impressions of the ad to a certain keyword based on the bid that is done. Their method, Generative Model based Ad Impression Forecasting Method, get better results in terms of accuracy when compared to Normalized Bid Model.

Chapter 3

Approach

This chapter describes the developed approach during the course of this dissertation. First we describe its high level architecture, and then we move on to describe each of the main phases that comprise it. For each phase, we describe the phase both at a general level and its materialization (its main goal, the most important variables, constraints, and methods).

Finally, we describe the experimental setup of the approach, and the setup for each component, and introduce the next chapter.

3.1 High-level Architecture

As the figure 3.1 makes clear, the goal of the proposed approach is to use a dataset containing logs of the web activity from an online advertising related network and use this information to generate a possible future web activity logs on the same network. This should be done in order to preserve tendencies and into data coherency.

This approach can be divided into three main phases:

1. *Segmentation*, which its main purpose is divide the dataset in smaller (by reducing the number of instances) and more predictable datasets, in order to improve the results obtained after the second phase, mostly when there are large quantities of data available.

2. The second phase is where the *forecast of the volumes* that characterize the traffic on the network are done, using time series prediction methods.

3. The third and last phase of the process, the more complex one, is where the volumes generated from the phase two combined with the data provided by the original dataset are used to *generate a dataset* that represent a possible future of the web activity on the target network.

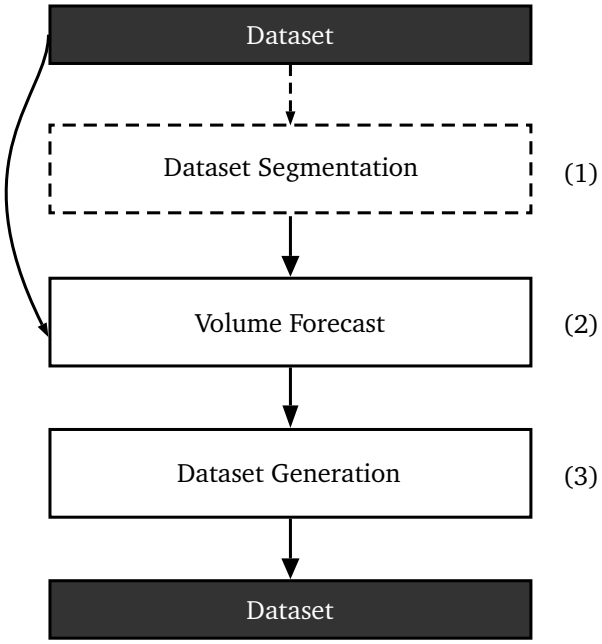

Figure 3.1: High level overview of the approach

3.2 Architecture for Web Activity Forecasting and Synthesising

3.2.1 Data Segmentation

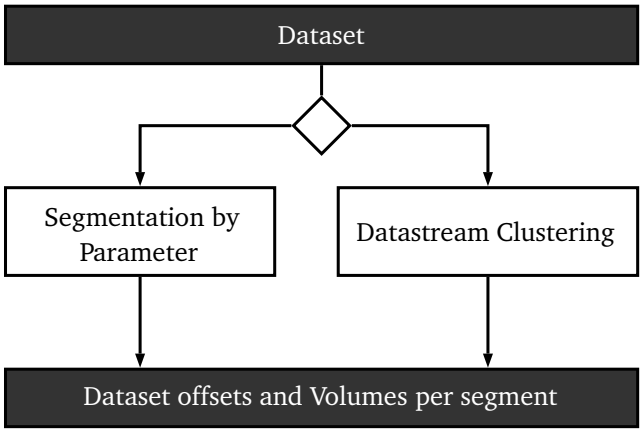

Figure 3.2: High level overview of Data Segmentation

The *Data Segmentation* phase was designed in order to achieve better results in the following phases.

In order to get better results in the second phase (Volumes Forecast) the dataset needs to possess certain characteristics that improve its predictability, or in other words, it needs to have recurring trends and /or patterns. One way of getting results is by splitting data

by parameters known to have an effect on traffic seasonality (for example the website that is being accessed).

As shown in figure 3.2 the method described before was one of the approaches chosen to achieve a more predictable time series. This method allows clustering of the dataset via a selected parameter, which returns *impression clusters* in which all members of a cluster share the same value in the parameter. This method requires great knowledge about the available impressions to be successful.

The other approach that was implemented in this phase is a simplified and striped down version of a *data stream clustering* algorithm. This algorithm is distance based, and since the structure of the datasets used on this thesis are not guaranteed, the distance measure used is naïve. The distance between impressions can be described as:

$$distance_{x,y} = \sum_{i=0}^{n} d_i$$

$$\text{where } d_i = \begin{cases} 1 \text{ if } x_i ! = y_i \\ 0 \text{ otherwise} \end{cases} \quad \text{and x, y are two impressions from the dataset}$$

Data: *lines*, all impressions available on the dataset; *threshold*, maximum distance to be considered member of a certain cluster.
Result: A list of clusters, each one characterized by the *medoid* (it is defined by and impression) and a list of offsets representing the position of each impression on the dataset.

```
(1) repeat
(2)     compare each impression with the existing list of clusters;
(3)     if dist < threshold then
(4)         add to the selected cluster;
(5)     else
(6)         create a new cluster and use the selected impression as the centroid for the
            new cluster;
(7)     end
(8) until no more impressions;
```

Algorithm 1: Data stream clustering simplified algorithm to aggregate the impressions by the parameters that they have in common.

This kind of algorithm was selected because, of the constraints imposed by the problem, the most important ones are the huge volume of data this approach needs to process, and the number and type of attributes that compose an impression which may vary from dataset to dataset.

The huge volume of data invalidates the usage of dissimilarity matrices due to memory constraints. The uncertainty of the parameters that are available also limit the quality of

the distance measure, since it assumes that every parameter has the same weight as the others.

This family of clustering algorithms has some known limitations, for example, the order in which the dataset is read directly influences the outcome of the algorithm. There is also a problem with the centroids, since they are represented by the attributes of the impression that originated that particular centroid, and as such are never updated, when a new impression has a distance lower than the threshold when compared with a certain centroid, it is not assured that the respective cluster is the most similar to the new impression. In order to address this possible issue, the ability of calculate the distance between the new impression and the available centroids and then select the one with least distance was also implemented.

3.2.2 Volume Forecasting

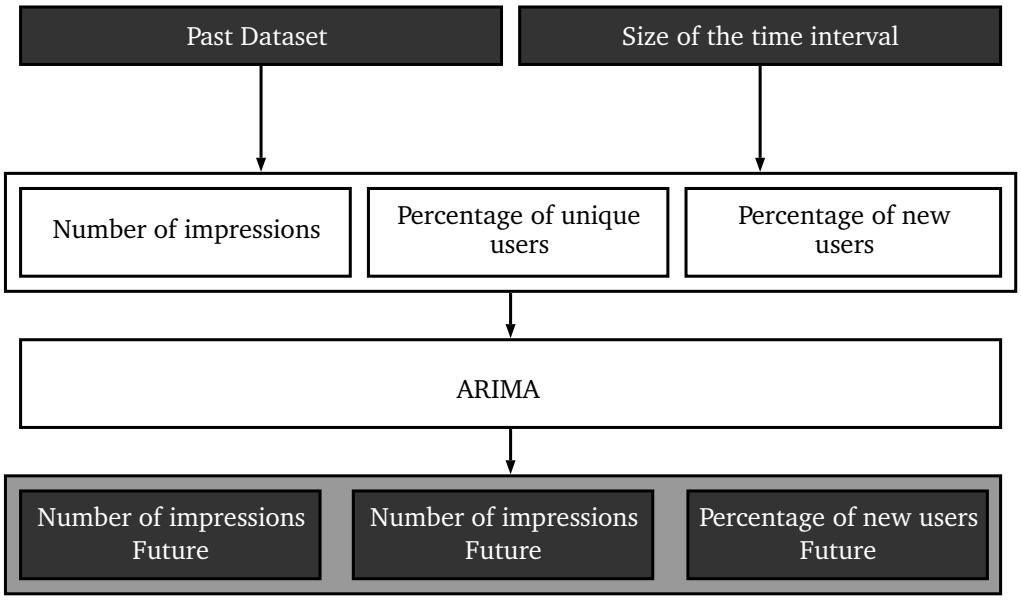

Figure 3.3: High level overview of Data Segmentation

One way to describe the impressions is by representing how many happen during a certain amount of time. So the impressions in the past can be described as time series of volume of impressions, composed by uniformly spaced time intervals.

So to predict how many impressions will happen in the future *time series analyses* techniques can be used.

Since the volume of impressions doesn't allow to fully characterize the future we need to use more time series of different variables which can provide more information about each epoch.

The approach here described involves the usage of three time series, one for the number of impressions, other for the percentage unique users and a last one for the percentage of unique users that had never appeared in the past.

Number of impressions is the number of entries on the dataset during each time interval.

Percentage of unique users is described by $\frac{\text{Number of users}}{\text{Number of impressions}} * 100$, and represents how the number of users are related with the number of impressions.

Percentage of new users represents per unit of time how many users that had never appeared before are represented. It is described by $\frac{\text{Number of new users}}{\text{Number of unique users}} * 100$.

So each unit of time can be represented by these three values. Now to characterize the future, we need to forecast of these three values on future time intervals. To complete this task the *ARIMA* approach was used separately for each of the three variables.

3.2.3 Dataset Generator

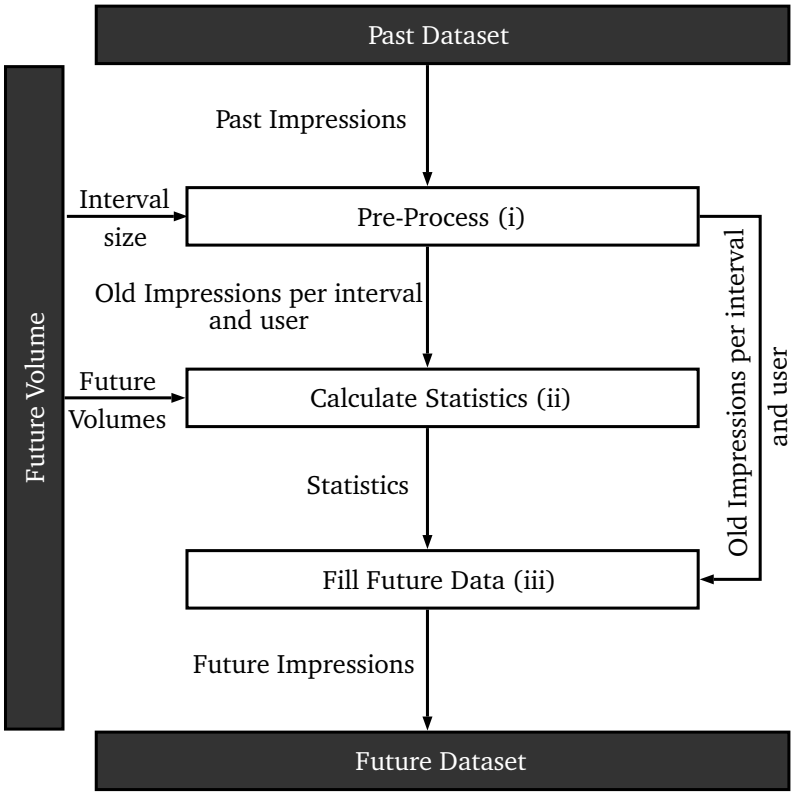

Figure 3.4: High level overview of the Dataset Generator

The last phase of the proposed approach generates a new dataset based on the original dataset and the values of the three time series generated on the previous phase (Volume Forecasting).

As shown in figure 3.4, this phase was divided into three sub-phases: the *pre-process* (3.2.3.1), *calculate statistics* (3.2.3.2), and *fill future data* (3.2.3.3).

The pre-process sub-phase (3.2.3.1), breaks the old dataset by interval size that was chosen on the volume forecasting phase (3.2.2) and organizes the impressions per interval and users.

The statistics calculation sub-phase, takes as input the processed data from the past dataset and the volumes from the volume prediction phase (3.2.2), verifies whether the values makes sense in the real world, for example the number of impressions is higher than the number of users (it is impossible for this type of dataset to have users with zero impressions, since the dataset represents impressions), break the impressions into smaller intervals, and use weekly historical data to predict the distribution of the interval volumes on the smaller intervals.

The last sub-phase, fill future data (3.2.3.3), is the most crucial phase of this process. It uses the volumes from the last sub-phase, and then selects past impressions using the restrictions imposed by the calculated volumes. This phase is also responsible for the generation of new users.

3.2.3.1 Pre-process (i)

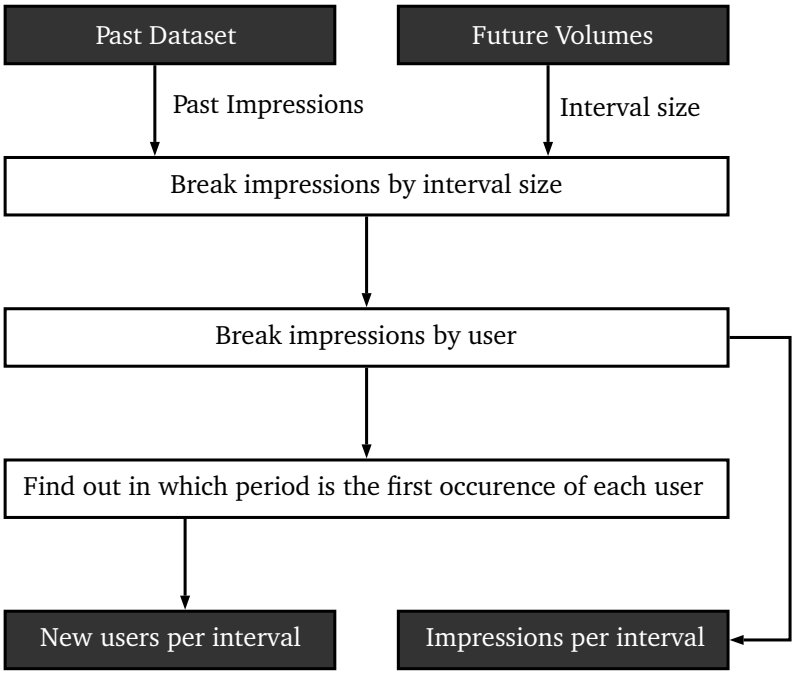

Figure 3.5: High level overview of the Dataset Pre-processing

This phase (figure 3.5) can be considered the simpler phase of the whole approach. Its main goal is to provide impressions organized by a defined time interval and place the first occurrence of each user in the correct interval, to better organize data for the next phases.

This is simply done by reading each line of the dataset and comparing with the start date of the dataset placing it on the correct interval. It also identifies which user is responsible for that impression and, if it is the first occurrence, place it as a new user for the interval where the impression belong.

3.2.3.2 Calculate Statistics (ii)

The main purpose of the sub-phase represented in the figure 3.6 is to distribute the volumes predicted for a certain interval in smaller periods based on the historical distribution of the volumes for the same period.

In order to obtain better resolution of how the data is distributed on the interval, these intervals are broken down in periods of one hour each. Now to understand how the volumes of the interval are distributed through the time periods, the data from the same interval up to two weeks before the interval which is currently being processed is used. For each hour period of historical interval the percentage of the total volume is calculated. To calculate the distribution , the input is the mean of the values from the same interval, the week before, and two weeks before. If there is only one previous week available, then the value of the current period is the same as the value from the week before; if there are no previous weeks available, the distribution of the interval volumes is done by dividing uniformly by each hour period. After the percentages are calculated, they are multiplied by the total volumes to obtain the absolute values for each interval.
For each hour period the following values are calculated:

- **Number of impressions**, this value represents how many requests were done during that space in time;
- **Number of users**, number of different users present during that space in time;
- **Number of first occurence users**, number of different users that occur for the first time in the interval during that specific hour period;
- **Number of new users**, number of different users that are completely new to the dataset and that occur for the first time in that specific hour period.

In order to be valid, these variables have to respect these constraints:

- the number of impressions must be equal or greater than the number of users for a specific period;

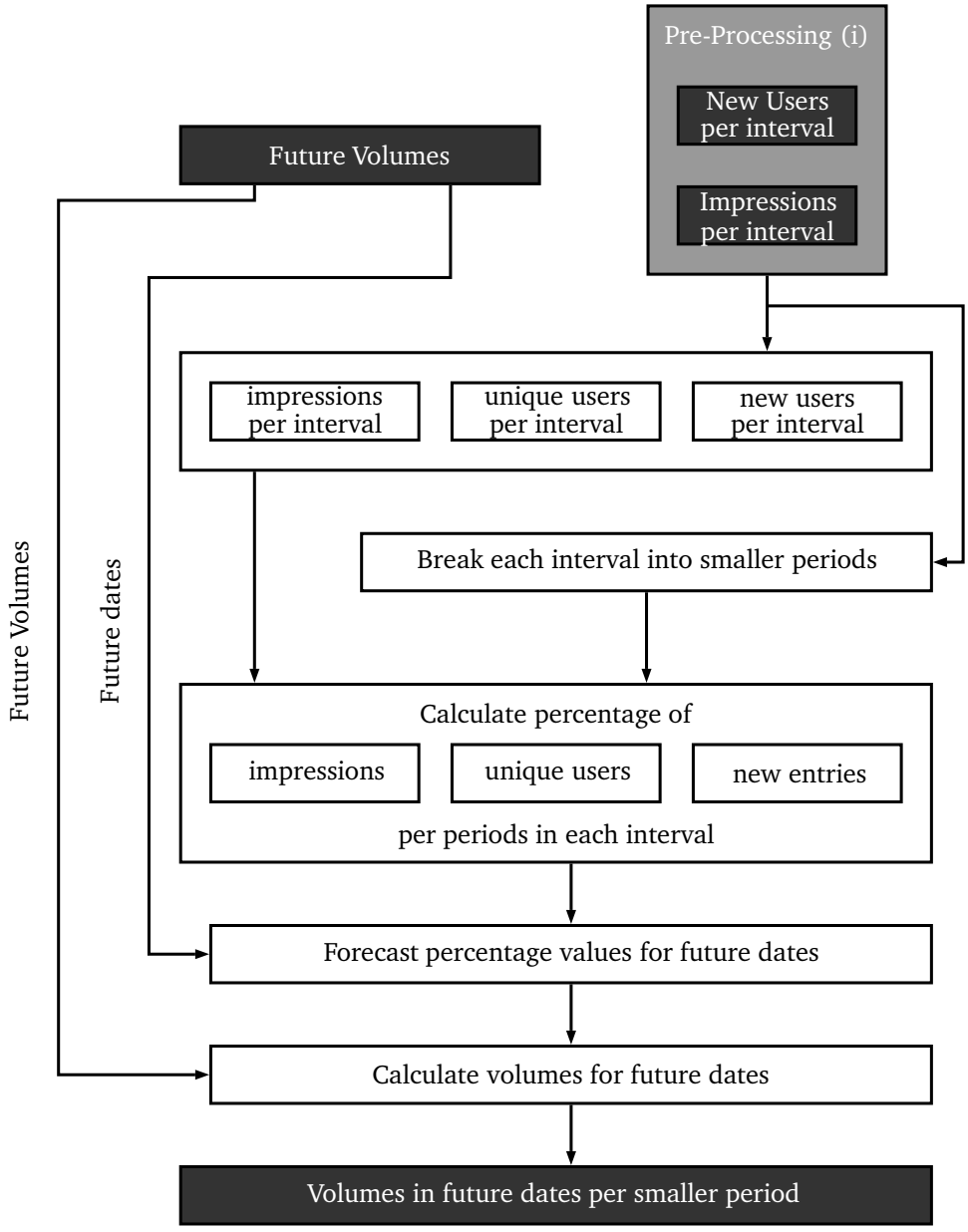

Figure 3.6: High level overview of statistics calculation

- the number of users must be less or equal to the sum of number of first occurrence users;
- the number of first occurrence users plus the number of new users must be equal or less than the number of users for the time period.

To guarantee the integrity of the result, these conditions have to be verified. So, after the calculation of the results for every hour period of a certain interval of time, a constraint check is performed and the results are adjusted accordingly. And to do so the approach

implemented according to algorithm 2 can be used to fulfil this task.

Data: *periods*, a list of one tuple for each period; *volumes*, a tuple containing the expected values for the interval
Result: A list of one tuple for each period, containing the adjusted values to meet the imposed constraints

```
(1) while the sum tuples for all periods != volumes do
(2)     diff = volumes - sum(periods);
(3)     for each element of periods do
(4)         To each constraint not respected, change the value according to the rule and
            then add the difference to the respective diff;
(5)         if any constraint not violated and the respective value of diff is not 0 then
(6)             change the respective value while still respect the imposed constraint;
(7)         end
(8)     end
(9) end
```

Algorithm 2: Values adjustment to meet the imposed constraints algorithm

At the end of this phase we have calculated all the four values for every future interval for which we want to predict the impressions.

3.2.3.3 Fill Future Data (iii)

The last sub-phase of the dataset generation phase is responsible for selecting which users and impressions, from the dataset that represents past activity, will reappear in the future with some changes.

To be able to capture the correct characteristics, so that the future follows the trends of the past, the first step must be the selection of the relevant periods of time in the past that are equivalent to the one which is currently being filled. This approach uses the same period from previous weeks ranked by proximity to the date, which makes more likely to use users from the previous week than two weeks ago.

After the correct dates are found, the users that will reappear in a given period must be selected taking into account the limits imposed by the volumes predicted on the previous sub-phase (3.2.3.2).

The process of the selection of old users is non-deterministic (uses random generated values to select which past users will reappear), but assures that does not enter an infinite loop, by terminating if it spends more than a *threshold* of cycles without getting a valid result. This is done in reverse chronological order, where we first select users from the week before. If there are not enough users we use data from previous weeks.

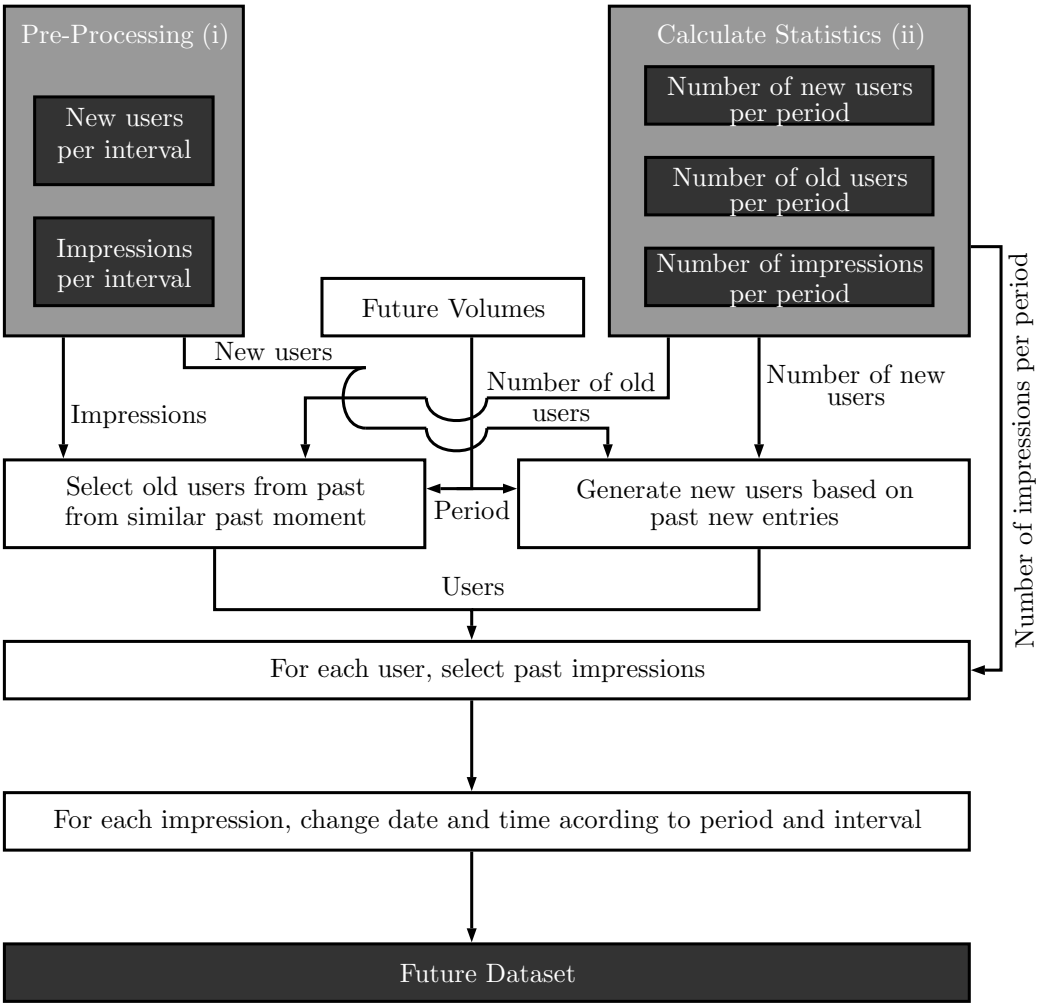

Figure 3.7: High level overview of fill future data

To complete the process of selection of users that will appear in the given period, we need to ”generate” new users. Since we have no knowledge of the data that represents an impression other than it has a date and an user identification token, this process is more a selection, than a generation, of past users that occurred for the first time on similar periods of the past, and then their date and user identification are changed according to the period for where this user will reappear. Like the previous selection process, this one also takes into account the chronological distance between events. First we use new occurrences from the same period, starting on the previous week until the first date of the past dataset (if there are not enough users available, users from the same interval[1] are used, if more users are needed then users from any interval of the dataset are used)(Algorithm 3).

To complete the process of filling the future impressions, we still need to select which

[1]Note that an interval is composed by multiple one hour periods.

Data: **users_period**, a list of lists of users reverse chronologically ordered;
users_from_same_interval, list of lists of users from the same interval as the one to be filled, also in reverse chronological order;
all_users, list of all users from the past;
period, date and time interval which the new users will be part of;
target_new_users, number of new users;
Result: A list of new users, containing the impressions with user identification token, modified and timestamp, adjusted to the given period.

(1) *selected_users* is initialized as an empty list;
(2) **for** *each* date *in the past, in reverse chronological order* **do**
(3) | **if** *length of* selected_users *less than* target_new_users **then**
(4) | | Randomly select users from *users_period* for the *date* that are not in *selected_users*, until there are no more users or the *target_new_users* is reached;
(5) | **else**
(6) | | break;
(7) | **end**
(8) **end**

(9) **if** *length of* selected_users *less than* target_new_users **then**
(10) | **for** *each* date *in the past, in reverse chronological order* **do**
(11) | | **if** *length of* selected_users *less than* target_new_users **then**
(12) | | | Randomly select users from *users_from_same_interval* for *date* that are not in *selected_users*, until there are no more users or the *target_new_users* is reached;
(13) | | **else**
(14) | | | break;
(15) | | **end**
(16) | **end**
(17) **end**

(18) **if** *length of* selected_users *less than* target_new_users **then**
(19) | **for** *each* date *in the past, in reverse chronological order* **do**
(20) | | **if** *length of* selected_users *less than* target_new_users **then**
(21) | | | Randomly select users from *all_users* for *date* that are not in *selected_users*, until there are no more users or the *target_new_users* is reached;
(22) | | **else**
(23) | | | break;
(24) | | **end**
(25) | **end**
(26) **end**

(27) For all *selected_users* replaces the UUID and timestamp with a randomly generated ones (timestamp belongs to given *period*);

Algorithm 3: User ”generation” process.

past impressions from the selected users will be used, and modify them according to the period that is being computed. First we randomly select one impression from each user, since all the selected users must appear at least one time. Afterwards, all the impressions are sorted in a random order, and the impressions are selected randomly one by one until the requirements are fulfilled. The selection process takes into account the number of impressions associated with a user; a user with many impressions in the past has more chances for one of his impressions to be selected. After all impressions are selected, their time value are updated to a random time inside the current time interval.

These steps must be executed for each period that we want to predict.

At the end we have a complete dataset that represent the volumes predicted on the Volume Forecasting phase (3.2.2), which are ready to be used on a simulator.

3.3 Experimental Setup

3.3.1 Dataset format

To develop this approach a dataset containing the time of the occurrence, user id token, and other information, like the browser, location, cookies, etc. was used, an example of this dataset in on the table 3.1.

Time	UserId	AdvertiserId	OrderId	LineItemId
2013-11-26-00:00:01	89e3b953-4422-49cb-bc10-8e869b30f0ab	26621901	138907941	107293701
CreativeId	**CreativeVersion**	**CreativeSize**	**AdUnitId**	**CustomTargeting**
32278541781	1	1x2	27191421	pos=0;showroom=ab
Domain	**CountryId**	**Country**	**RegionId**	**Region**
2e07dc054c5bdcec109605689ec8e11f	2276	Germany	20240	Saxony-Anhalt
CityId	**City**	**BrowserId**	**Browser**	**OSId**
1004957	Hettstedt	500072	Google Chrome	501011
OS	**OSVersion**	**BandWidth**	**TimeUsec**	**AudienceSegmentIds**
Microsoft Windows 7		adsl-8mbps	1384817822	
Product	**RequestedAdUnitSizes**	**BandwidthGroupId**		
Ad Server	1x2	3		

Table 3.1: Example data from the dataset used with the respective label

The only fields that the solution depends on are the time and the user id token, all the other fields are optional and the approach described above does not have any specific knowledge about them, nor it depends on it. The optional parameters can be useful on the segmentation phase, to divide the dataset into smaller datasets in order to capture more specific characteristics.

The design decision of not using anything besides the time and user id, was made because the proposed solution should be agnostic of the data that it analyses. Mostly because the data available is not standardised, and the solution should be able to be used on data from multiple sources, with each source having its own way of saving activities.

3.3.2 Experimental Setup configurations

In order to be able to test the previously described approach, the proposed approach was implemented using *Python*. The experimental setup is completely modular, with every phase being totally independent of the other. In other words, any phase can use as input data from other sources other than the implemented ones. Other than *Python*, *R* was used in order to use the *ARIMA* model.

Segmentation

In this phase there are two implement methods, as explained before in section 3.2.1. The $datastream - based$ algorithm has a single configuration parameter, the threshold for the distance between the impression that defines the clusters and the one that is being tested. In the case of segmentation by parameters, the parameters to be used for the division can be changed, so different configurations were tested.

This phase cannot be directly validated nor compared, so the results were only validated and compared after the forecast phase.

The baseline method for this phase, divides the impressions on a chronological order into 100 groups.

Volume Forecast

On this phase, the *ARIMA* model is used to predict the behaviour of the time series for the time interval where we want to predict the volume values. To use the *ARIMA* model, the function $auto.arima$ from the package $forecast$ [HK07] (from *R*) was used, using a *Python* wrapper in order to process the data. The *R* is only responsible for the prediction.

The function used allows the configuration of various parameters, most of which were covered by the tests (for example, the usage of *drift terms*). The frequency of the time series object in *R* is also configurable.

Different time intervals were also tested, in order to understand which interval produces better results.

To compare these results, *RMSE* and *MASE* where used.

The baseline method for this phase, copies data from the furthest date of the dataset into the starting date of the forecast, respecting the day of the week and the hour.

Dataset Generation

This last phase does not take any additional parameters. It uses the values that had been forecast in the volume forecast phase and the old dataset in order to build a dataset which represents the future.

This phase was tested using synthetic generated past datasets which contained certain changes in some parameters (examples: changes in browser usage and addition of new domains).

Since a simulator was not available during the test phase, some queries over the data were done to test the capability of this phase.

3.4 Conclusion

As mentioned before, this process has multiple phases where it tries to capture the maximum amount of information about the past in order to more accurately predict the future. This chapter ends with an high level overview of the experimental setup, for which the results will be presented and analysed on the next chapter.

Chapter 4

Results and Analyses

In this chapter, we present some results of experiments done using the approach developed (Chapter 3) and with multiple datasets (with different characteristics).

We also present some experiments, made in order to understand which values are better for some parameters, and in which situations they work better or worst.

4.1 Interval size without segmentation

To test which intervals of time get better results for predicting time series used on the approach, a dataset containing real data from 26-09-2013 to 27-01-2014 was used. The interval sizes tested were: 4 hours, 6 hours, 8 hours, 12 hours and 24 hours (arbitrary choosing).

For this test only the *Volume Forecasting*(Section 3.2.2) phase, was used.

The time intervals used for the test were:

- case 1 - The results presented on table 4.1 use 26-09-2013 to 26-11-2013 for training; 27-11-2013 to 25-01-2014 for validation;
- case 2 - The results presented on table 4.2 use 26-09-2013 to 26-12-2013 for training; 27-12-2013 to 25-01-2014 for validation;
- case 3 - The results presented on table 4.3 use 26-09-2013 to 31-12-2013 for training; 01-01-2014 to 25-01-2014 for validation;

For each one of the three cases the error of the impressions volume forecasting for every test interval is shown. In every case, the interval that got the best results was the 12 hour interval (minimum *MASE*).

Additional errors values and graphs for this test are available on Appendix A (for case 1), Appendix B (for case 2) and Appendix C (for case 3).

	4h baseline	4h allow drift	4h	6h baseline	6h allow drift	6h
σ (Real Data)		946.25			1358.24	
RMSE	690.87	562.80	562.80	945.70	742.57	742.57
MASE	0.5227	0.4397	0.4397	0.3659	0.3003	0.3003

	8h baseline	8h allow drift	8h	12h baseline	12h allow drift	12h
σ (Real Data)		1785.84			2224.74	
RMSE	1150.43	955.35	955.35	1555.01	1233.82	1233.82
MASE	0.2712	0.2473	0.2473	0.2253	0.1967	0.1967

	24h baseline	24h allow drift	24h
σ (Real Data)		1946.70	
RMSE	2448.97	2139.99	2139.99
MASE	1.4338	1.2725	1.2725

Table 4.1: Case 1: Forecast errors for different interval sizes (best result in red)

	4h baseline	4h allow drift	4h	6h baseline	6h allow drift	6h
σ (Real Data)		908.70			1293.78	
RMSE	727.31	1083.22	834.32	967.46	740.73	740.73
MASE	0.1922	0.3152	0.2348	0.1312	0.0947	0.0947

	8h baseline	8h allow drift	8h	12h baseline	12h allow drift	12h
σ (Real Data)		1691.32			2154.85	
RMSE	1205.49	999.47	999.47	1565.43	1722.83	1722.83
MASE	0.0983	0.0905	0.0905	0.0857	0.0953	0.0953

	24h baseline	24h allow drift	24h
σ (Real Data)		1752.65	
RMSE	2329.76	1646.92	1646.92
MASE	0.4198	0.2925	0.2925

Table 4.2: Case 2: Forecast errors for different interval sizes (best result in red)

	4h baseline	4h allow drift	4h	6h baseline	6h allow drift	6h
σ (Real Data)		929.73			1334.01	
RMSE	737.22	692.61	692.61	985.86	930.13	930.13
MASE	0.1548	0.1441	0.1441	0.1071	0.0998	0.0998

	8h baseline	8h allow drift	8h	12h baseline	12h allow drift	12h
σ (Real Data)		1738.75			2203.12	
RMSE	1197.45	1165.13	1165.13	1581.66	1453.84	1453.84
MASE	0.0778	0.0763	0.0763	0.0695	0.0595	0.0595

	24h baseline	24h allow drift	24h
σ (Real Data)		1686.92	
RMSE	2275.91	2547.73	2260.99
MASE	0.3304	0.3525	0.3044

Table 4.3: Case 3: Forecast errors for different interval sizes (best result in red)

4.2 Segmentation

The experiments described on this section use all the phases of the approach. This experiments used datasets generated on an artificial environment, in order to test the ability of the approach to capture changes on the volumes of some characteristics of the dataset. We also used a real world dataset, in order to assess the result in more realist conditions.

4.2.1 Particular browser increasing

Without segmentation

For the first test an artificial generated dataset was used, containing a constant overall volume of impressions, but with the particularity of an increasing volume of impressions from users using ”Safari 4.0” browser.

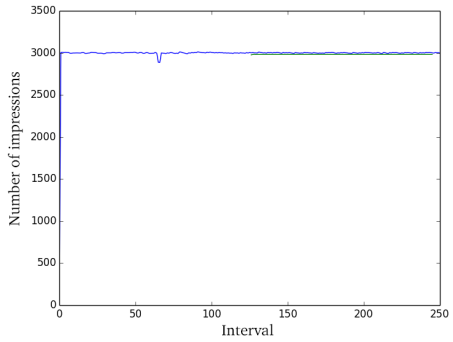

Figure 4.1: Volume impression forecast, using 12h period without clustering (blue: real; green: forecast)

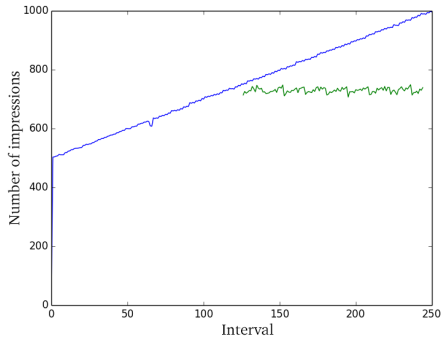

Figure 4.2: Volume impression forecast, using 12h period without clustering, filtered by "Safari 4.0" (blue: real; green: forecast)

σ (Real Data)	RMSE	MASE
3.33	19.61	0.6760

Table 4.4: Error for impression volume forecast, using a 12h period without clustering

σ (Real Data)	RMSE	MASE
72.28	155.42	19.5770

Table 4.5: Error for impression volume forecast, using a 12h period without clustering, query for ”Safari 4.0”

As shown in the figure 4.1 and in the table 4.4, (without the segmentation phase) the proposed approach was able to forecast the total volume of impressions only with a small error.

In order to be able to see the result of the approach for the behavior of the ”Safari 4.0” users we need to generate a future dataset (last phase of the approach). The result for the query for the ”Safari 4.0” browser, over the resultant dataset, can be seen on figure 4.2.

As shown on the figure 4.2, the dataset generation phased mimics the behavior of that particular characteristic from the previous week into the future, which in this case is not a good result.

Clustering Baseline

In order to better compare the results of the different segmentation approaches, a baseline method was used[1].

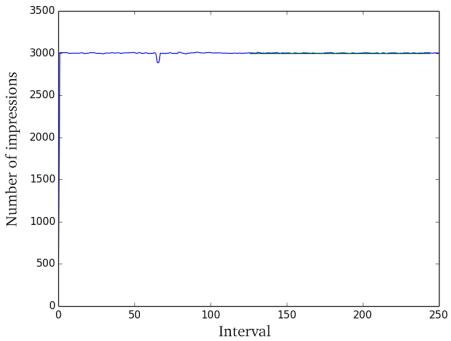

Figure 4.3: Volume impression forecast, using 12h period with baseline clustering (blue: real; green: forecast)

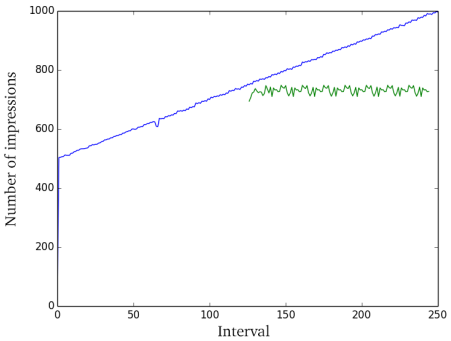

Figure 4.4: Volume impression forecast, using 12h period with baseline clustering, filtered by "Safari 4.0" (blue: real; green: forecast)

σ (Real Data)	RMSE	MASE
3.33	3.38	0.0931

Table 4.6: Error for impression volume forecast, using 12h period with baseline clustering

σ (Real Data)	RMSE	MASE
72.28	154.99	19.5144

Table 4.7: Error for impression volume forecast, using 12h period with baseline clustering, filtered by "Safari 4.0"

On figure 4.3 and table 4.7 we can assess that this approach gets a slight improvement over the non-segmented approach. If we look at the filtered data (Figure 4.4) the improvement is not that noticeable.

Segmentation by Parameter

Segmentation by browser

Since we know that this particular dataset has a peculiarity, as previously seen on filtered results (for example Figure 4.6), in the browser parameter impression volume. To verify how the prediction would react to the clustering by parameter(specifically by the browser

[1] the original dataset was grouped in a chronological order on 100 different groups.

attribute) this will forecast the volumes values for each browser present on the dataset. This allow us to get better results in predicting the traffic volumes for each browser.

As we can see on Figure 4.5, the overall error (Table 4.8) for impression volume forecasting is worse than without the clustering phase. This is due to the fact that the prediction error associated with each cluster is amplified when the volume data is joined back together.

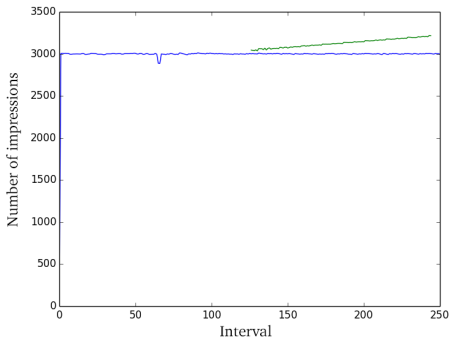

Figure 4.5: Volume impression forecast, using 12h period with clustering by the browser attribute (blue: real; green: forecast)

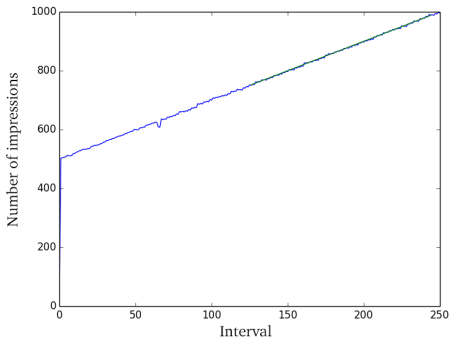

Figure 4.6: Volume impression forecast, using 12h period with clustering by the browser parameter, filtered by "Safari 4.0" (blue: real; green: forecast)

σ (Real Data)	RMSE	MASE
3.33	136.78	4.4253

Table 4.8: Error for impression volume forecast, using a 12h period with clustering by browser attribute

σ (Real Data)	RMSE	MASE
72.28	2.58	0.2892

Table 4.9: Error for impression volume forecast, using a 12h period with clustering by browser attribute, query for ”Safari 4.0”

In the other hand, in the Figure 4.6(which show the results filtered by ”Safari 4.0” browser) we get a better result prediction for this characteristic, as shown by the error values on Table 4.9. This tells us that *clustering by parameter* gives better predictions for certain characteristics, by capturing the volume for each one in the past and propagating it into the future.

Using datastream-based clustering

Using the datastream-based clustering method, with a *threshold* of 14 (which means that each group member of each groups will only have a maximum of 14 parameters different than the centroid of the group) the result for the overall impressions volume was worse than methods without clustering and the baseline clustering method, but better than the clustering by browser. If we analyse the figure 4.8, we can assess than the result is only beat by the clustering for that specific parameter, which ultimately is a good result.

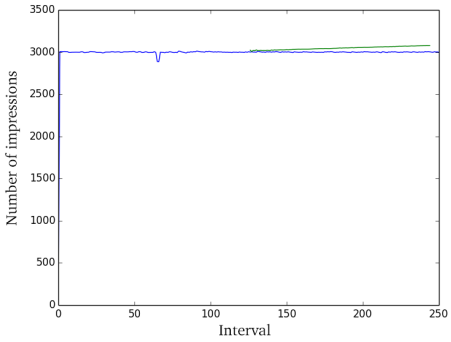

Figure 4.7: Volume impression forecast, using 12h period with clustering based datastream-based clustering, with a *threshold* of 14 maximum distance (blue: real; green: forecast)

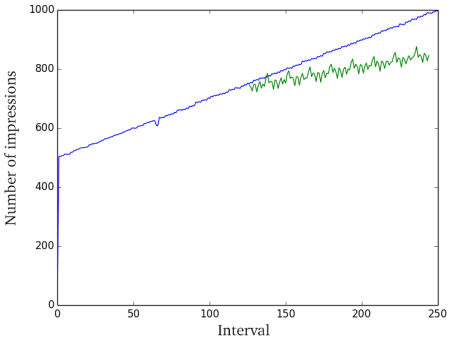

Figure 4.8: Volume impression forecast, using 12h period with clustering based datastream-based clustering, with a *threshold* of 14 maximum distance, filtered by "Safari 4.0" (blue: real; green: forecast)

σ (Real Data)	RMSE	MASE
3.33	50.41	1.6226

Table 4.10: Error for impression volume forecast, using 12h period with clustering based datastream-based clustering, with a *threshold* of 14 maximum distance

σ (Real Data)	RMSE	MASE
72.28	84.78	10.5275

Table 4.11: Error for impression volume forecast, using 12h period with clustering based datastream-based clustering, with a *threshold* of 14 maximum distance, filtered by "Safari 4.0"

4.2.2 Specific Domain decreasing linearly

For this set of tests another artificially generated dataset was used, in this particular case one of the domains represented on the dataset, with a steady linear decrease following a linear function.

Without Segmentation

σ (Real Data)	RMSE	MASE
4.33	5.00	0.1392

Table 4.12: Error for impression volume forecast, using 12h period without segmentation

σ (Real Data)	RMSE	MASE
72.50	152.78	12.8542

Table 4.13: Error for impression volume forecast, using 12h period without segmentation, filtered by the domain ”ef4e08fb71d96d19406663f8bb7ce6c0”

Without any segmentation phase we obtained a good result for the overall impression volume prediction. If we filter the results for the particular domain that we know will be

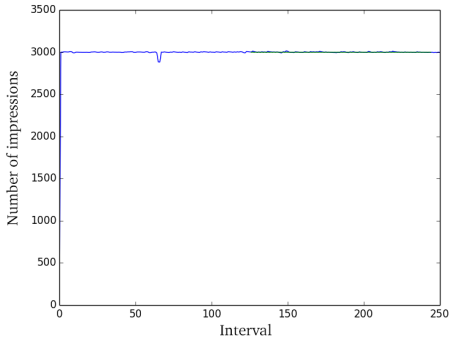

Figure 4.9: Impression volume forecast, using 12h period without segmentation (blue: real; green: forecast)

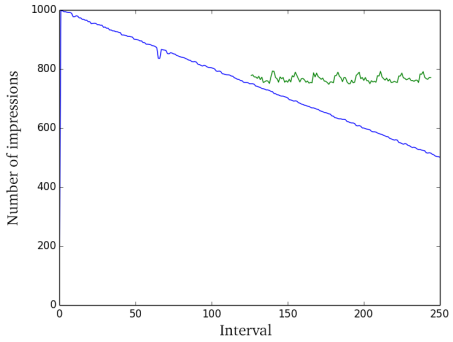

Figure 4.10: Impression volume forecast, using 12h period without segmentation, filtered by the domain ”ef4e08fb71d96d19406663f8bb7ce6c0” (blue: real; green: forecast)

decreasing, the result mimics the behaviour of the previous weeks since the information about that particular domain is limited.

Clustering Baseline

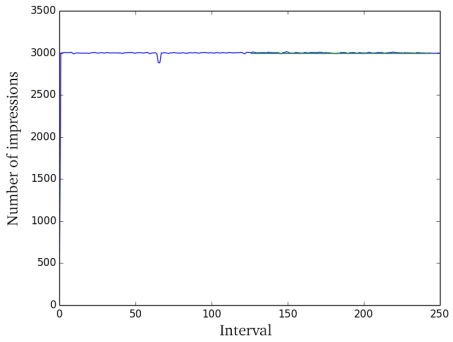

Figure 4.11: Impression volume forecast, using 12h period using baseline segmentation (blue: real; green: forecast)

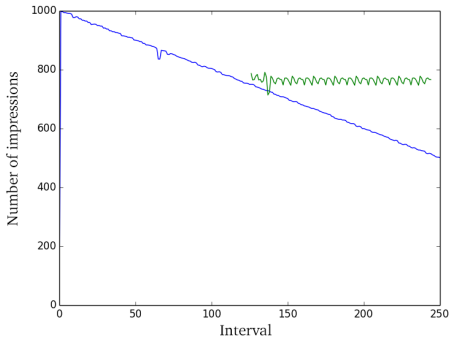

Figure 4.12: Impression volume forecast, using 12h period using baseline segmentation, filtered by the domain ”ef4e08fb71d96d19406663f8bb7ce6c0” (blue: real; green: forecast)

In the figure 4.11 and table 4.14 we can take a look at the results given by the segmentation baseline method for the total volume of impressions forecast for this dataset. The figures show the results are better than without segmentation, and when we filter the data for the domain that is decreasing in volume of impressions (figure 4.12 and table 4.15) we also obtain a slightly better result.

σ (Real Data)	RMSE	MASE
4.33	4.30	0.1176

Table 4.14: Error for impression volume forecast, using 12h period with baseline segmentation

σ (Real Data)	RMSE	MASE
72.50	150.7442	12.6627

Table 4.15: Error for impression volume forecast, using 12h period with baseline segmentation, filtered by the domain "ef4e08fb71d96d19406663f8bb7ce6c0"

Segmentation by Parameter

Segmentation by Domain

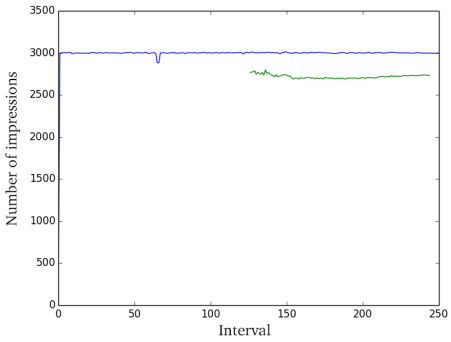

Figure 4.13: Impression volume forecast, using 12h period using segmentation by domain (blue: real; green: forecast)

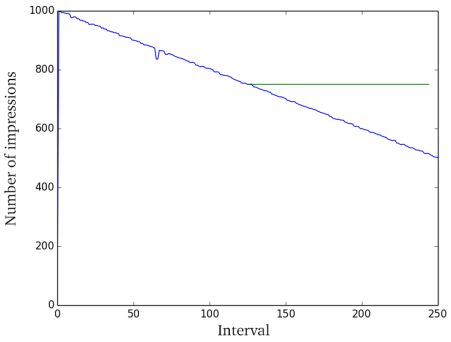

Figure 4.14: Impression volume forecast, using 12h period using segmentation by domain, filtered by the domain "ef4e08fb71d96d19406663f8bb7ce6c0" (blue: real; green: forecast)

σ (Real Data)	RMSE	MASE
4.33	283.20	9.9304

Table 4.16: Error for impression volume forecast, using 12h period with segmentation by domain

σ (Real Data)	RMSE	MASE
72.50	137.93	11.2913

Table 4.17: Error for impression volume forecast, using 12h period with segmentation by domain, filtered by the domain "ef4e08fb71d96d19406663f8bb7ce6c0"

Like we did for the previous test group (where we knew that a particular browser was increasing its value in volume impressions and so we segmented the data by browser), in this case we know that a domain is decreasing, so we tested segment the data by domain. In the figure 4.14 the results were not as expected. Even though we got a better result in the filtered data, when compared to the previous methods, the result for the total volume is much worse. This is probably due to the small volumes in each domain, since in this dataset there was about 4300 domains in around 3000 impressions for each 12h period, which can result in a very difficult to predict time series.

Using datastream-based clustering

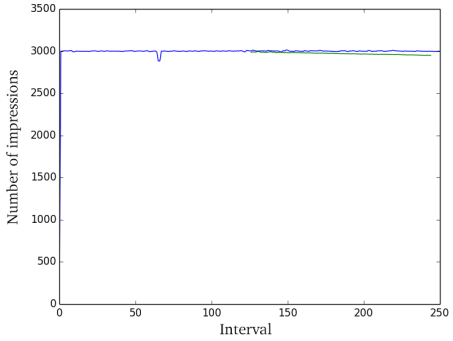

Figure 4.15: Impression volume forecast, using 12h period using datastream-based clustering with a *threshold* of 14 (blue: real; green: forecast)

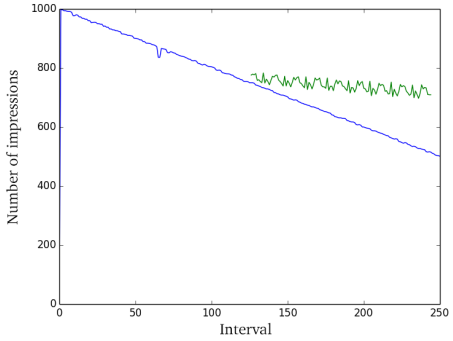

Figure 4.16: Impression volume forecast, using 12h period using datastream-based clustering with a *threshold* of 14, filtered by the domain "ef4e08fb71d96d19406663f8bb7ce6c0" (blue: real; green: forecast)

σ (Real Data)	RMSE	MASE
4.33	32.48	1.0613

Table 4.18: Error for impression volume forecast, using 12h period using datastream-based clustering with a *threshold* of 14

σ (Real Data)	RMSE	MASE
72.50	123.27	10.3919

Table 4.19: Error for impression volume forecast, using 12h period using datastream-based clustering with a *threshold* of 14, filtered by the domain "ef4e08fb71d96d19406663f8bb7ce6c0"

The datastream-based segmentation approach using a *threshold* equal to 14 of maximum distance, we get a slightly worse result than the method without segmentation and the baseline segmentation approach, as shown by Figure 4.15 and Table 4.18. This is the method responsible for the best results on this series, in terms of capturing the peculiarity of this dataset, as shown by figure 4.16 and table 4.19.

4.2.3 Real Data

For this last series of tests we used a dataset containing real data. Since we don't know any particular characteristic to search for, we will compare the results for the query for "Portugal" as country of origin for each impression.

Without Segmentation

Like in the other datasets we start by testing the result without any segmentation. In overall volume of impressions we get a good result (figure 4.17 and table 4.20), and when

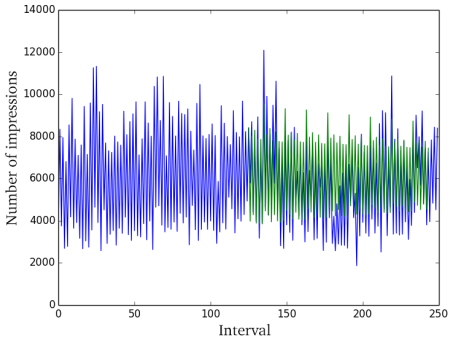

Figure 4.17: Impression volume forecast, using 12h period without clustering (blue: real; green: forecast)

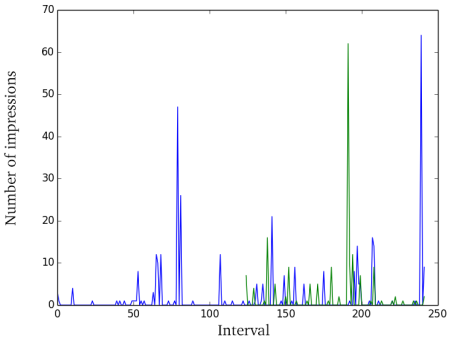

Figure 4.18: Impression volume forecast, using 12h period without clustering, filtered by country = Portugal (blue: real; green: forecast)

σ (Real Data)	RMSE	MASE
2224.74	1237.99	0.1958

Table 4.20: Error for impression volume forecast, using 12h period without clustering

σ (Real Data)	RMSE	MASE
6.70	9.24	1.2511

Table 4.21: Error for impression volume forecast, using 12h period without clustering, filtered by country = Portugal

over the result a query for origin country ”Portugal” is done, the result mimics the past characteristics and the result on this case is not bad.

Clustering Baseline

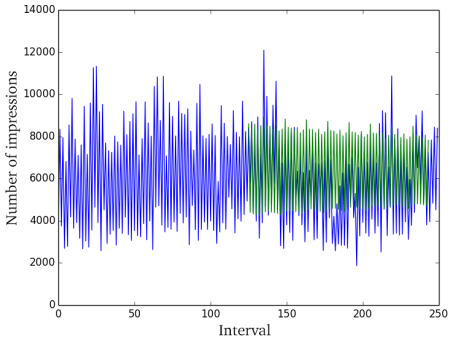

Figure 4.19: Impression Volume forecast, using 12h period with baseline segmentation (blue: real; green: forecast)

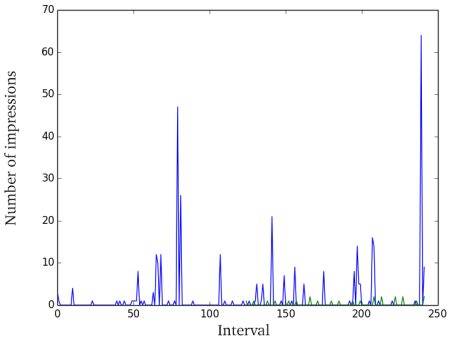

Figure 4.20: Impression Volume forecast, using 12h period with baseline segmentation, filtered by country = Portugal (blue: real; green: forecast)

In this case the baseline segmentation gets a worse result than the non-segmented approach in terms of total volume, but a better approximation in terms of the results for

σ (Real Data)	RMSE	MASE
2224.74	1352.42	0.2174

Table 4.22: Error for volume impression forecast, using 12h period with baseline clustering

σ (Real Data)	RMSE	MASE
6.70	6.88	0.7896

Table 4.23: Error for impression volume forecast, using 12h period with baseline clustering, filtered by country = Portugal

the query ”Portugal”. This is probably due to additional granularity gained by the data division.

Segmentation by Parameter

Segmentation by browser

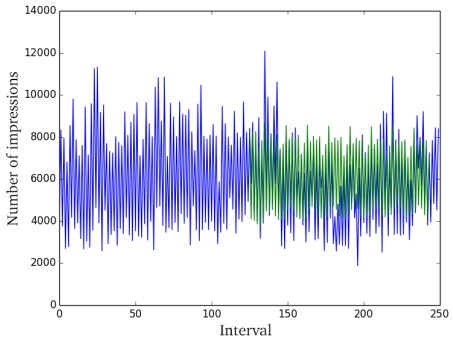

Figure 4.21: Impression Volume forecast, using 12h period with segmentation by browser(blue: real; green: forecast)

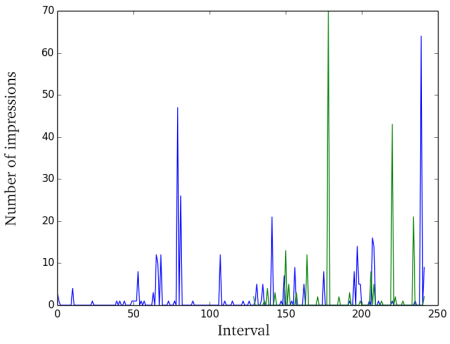

Figure 4.22: Impression Volume forecast, using 12h period with segmentation by browser, filtered by country = Portugal (blue: real; green: forecast)

σ (Real Data)	RMSE	MASE
2224.74	1226.60	0.1884

Table 4.24: Error for impression volume forecast, using 12h period with segmentation by browser

σ (Real Data)	RMSE	MASE
6.70	10.75	1.3883

Table 4.25: Error for impression volume forecast, using 12h period with segmentation by browser, filtered by country = Portugal

Since in the case of this dataset we do not know for which particular parameter we might want to segment by, the browser parameter was randomly selected as the segmentation parameter. For the overall volume result this was the best result over the other approaches over this particular dataset. In terms of the query for ”Portugal” the result was slightly worse than the previous approaches.

Using datastream-based clustering

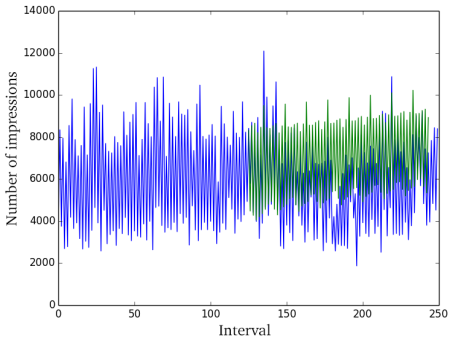

Figure 4.23: Impression Volume forecast, using 12h period with datastream-based segmentation using *threshold* of 20 (blue: real; green: forecast)

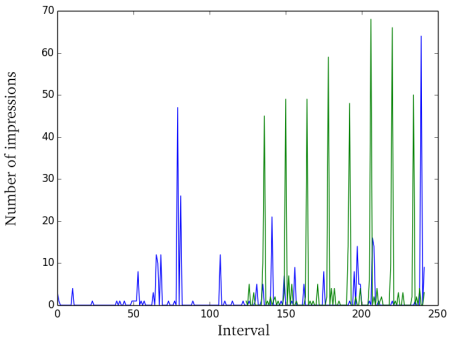

Figure 4.24: Volume impression forecast, using 12h period with datastream-based segmentation using *threshold* of 20, filtered by country = Portugal (blue: real; green: forecast)

σ (Real Data)	RMSE	MASE
2224.74	1786.43	0.2927

Table 4.26: Error for impression volume forecast, using 12h period with datastream-based segmentation using *threshold* of 20

σ (Real Data)	RMSE	MASE
6.70	15.93	2.5960

Table 4.27: Error for impression volume forecast, using 12h period with datastream-based segmentation using *threshold* of 20, filtered by country = Portugal

For this dataset the *threshold* for the distance needed to be increased to 20 (out of 28) because with a smaller *threshold* we would get too many groups, and some of them with data in the future were not represented on the training dataset.

In overall analysis this was the worst result for this dataset for impression volume forecast and for the queried data. This bad result is due to the fact that the results did not have any particular characteristic in terms of volume trend that made sense together, but were group because they had similar parameters.

4.3 Conclusion

In about every test case, the results that used the segmentation phase got better results.

These results were only compared to a small number of queries, so the results can be completely different when used in its goal environment, the online advertising campaigns impact on that particular network.

In order to be able to assess this impact a simulator should have been used to test the results of this approach at its fully capability. But due to time constraints this was not possible.

Chapter 5

Conclusions and Future Work

5.1 Objectives Fulfilment

We proposed the use of the three phase approach (Chapter 3) in order to obtain a richer future prediction of web activity on a network. The first phase allows better predictions for certain characteristics, the second phase allows the prediction of volumes in the future, and lastly the third phase fills the predicted future data with additional information from the past while following the trends of the predicted volumes.

Our results show that the usage of dataset segmentation returns good to moderate improvements while predicting the volumes.

At the end we get a dataset with the same format as the one given as input, but with future data that follows the main characteristics of the past while amplifying some of the most important characteristics for the future.

This dataset is completely ready to be used on simulations of online advertising campaigns in order to perceive their behaviour in the future.

5.2 Future Work

In order to achieve better results, some other approaches could be explored. For example, we could use *MLP*, multilayer perceptron, to predict the time series values instead of the *ARIMA*.

It would also make sense to explore the usage of additional time series for new entries of each parameter, for example in order to correctly predict the occurrence of a new website. This phase will also need additional knowledge about some of the parameters to get better results (for example know how to generate a new domain). This knowledge should be optional in order to maintain the data agnostic approach, but when available it would probably would improve the results.

Another road to be explored is the segmentation phase for example the usage recursive parameter segmentation. For example, segment the whole dataset by browser and then by country. It is my believe that this approach would get good results if a huge volume of data were available.

It would also be interesting to test the resultant datasets on a real online advertising campaign simulator in order to better assess the results in practical applications.

References

[AC92] J.Scott Armstrong and Fred Collopy. Error measures for generalizing about forecasting methods: Empirical comparisons. *International Journal of Forecasting*, 8(1):69 – 80, 1992.

[Adf] Adfonic. Adfonic's global admetrics report for q3 2013. http://adfonic.com/wp-content/uploads/2012/03/Adfonic_Global_AdMetrics_Q3_2013.pdf. Accessed last time on 10 Fev 2014.

[And04] Mangàni Andrea. Online advertising: Pay-per-view versus pay-per-click. *Journal of Revenue & Pricing Management*, 2(4):295 – 302, 2004.

[Ani10] A. Anitha. A new web usage mining approach for next page access prediction, 2010.

[Arm01] J.Scott Armstrong. Evaluating forecasting methods. In J.Scott Armstrong, editor, *Principles of Forecasting*, volume 30 of *International Series in Operations Research Management Science*, pages 443–472. Springer US, 2001.

[BA97] LEONARD A. BRESLOW and DAVID W. AHA. Simplifying decision trees: A survey. *The Knowledge Engineering Review*, 12:1–40, 1 1997.

[BBR+07] Michael Bensch, Dominik Brugger, Wolfgang Rosenstiel, Martin Bogdan, Wilhelm Spruth, and Peter Baeuerle. Self-learning prediction system for optimisation of workload management in a mainframe operating system, 2007.

[BC03] Kristin P. Bennett and Colin Campbell. Support vector machines: Hype or hallelujah? *SIGKDD Explorations*, 2:2000, 2003.

[BFGN14] N. Buchbinder, M. Feldman, A. Ghosh, and J. Naor. Frequency capping in online advertising. *Journal of Scheduling*, pages 1–14, 2014. Article in Press.

[BJR13] George EP Box, Gwilym M Jenkins, and Gregory C Reinsel. *Time series analysis: forecasting and control*. John Wiley & Sons, 2013.

[Bre01] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[BS04] R. Baraglia and F. Silvestri. An online recommender system for large web sites. In *Web Intelligence, 2004. WI 2004. Proceedings. IEEE/WIC/ACM International Conference on*, pages 199–205, Sept 2004.

[BV10] Nitin Bhatia and Vandana. Survey of nearest neighbor techniques. *CoRR*, abs/1007.0085, 2010.

[CML13] Olivier Chapelle, Eren Manavoglu Microsoft, and Romer Rosales Linkedin. A simple and scalable response prediction for display advertising. 2013.

[CS02] Koby Crammer and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *J. Mach. Learn. Res.*, 2:265–292, March 2002.

[CV95] Corinna Cortes and Vladimir Vapnik. Support-vector networks. In *Machine Learning*, pages 273–297, 1995.

[FFPs+96] Usama Fayyad, Usama Fayyad, Gregory Piatetsky-shapiro, Gregory Piatetsky-shapiro, Padhraic Smyth, and Padhraic Smyth. Knowledge discovery and data mining: Towards a unifying framework. pages 82–88. AAAI Press, 1996.

[FMK03] Enrique Frías-Martínez and Vijay Karamcheti. A customizable behavior model for temporal prediction of web user sequences. In OsmarR. Zaïane, Jaideep Srivastava, Myra Spiliopoulou, and Brij Masand, editors, *WEBKDD 2002 - Mining Web Data for Discovering Usage Patterns and Profiles*, volume 2703 of *Lecture Notes in Computer Science*, pages 66–85. Springer Berlin Heidelberg, 2003.

[Ger12] J. Gern. A guide to digital advertising: Making sense of the ad networks, management tools, and ad-serving solutions. *EContent*, 35(7):30–31, 2012.

[GH03] Mathias Géry and Hatem Haddad. Evaluation of web usage mining approaches for user's next request prediction. In *Proceedings of the 5th ACM International Workshop on Web Information and Data Management*, WIDM '03, pages 74–81, New York, NY, USA, 2003. ACM.

[GL99] Paul Goodwin and Richard Lawton. On the asymmetry of the symmetric {MAPE}. *International Journal of Forecasting*, 15(4):405 – 408, 1999.

[HK06a] Jiawei Han and Micheline Kamber. *Data Mining. Concepts and Techniques*. Morgan Kaufmann, 2nd ed. edition, 2006.

[HK06b] Rob J. Hyndman and Anne B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679 – 688, 2006.

[HK07] Rob J Hyndman and Yeasmin Khandakar. Automatic time series for forecasting: the forecast package for r. Technical report, Monash University, Department of Econometrics and Business Statistics, 2007.

[HR76] Laurent Hyafil and Ronald L. Rivest. Constructing optimal binary decision trees is np-complete. *Information Processing Letters*, 5(1):15 – 17, 1976.

[Iha98] R Ihaka. R: past and future history. 1998.

[JL07] Nien-Yi Jan and Nancy P. Lin. Web user behaviors prediction system using trend similarity. In *Proceedings of the 7th WSEAS International Conference on Simulation, Modelling and Optimization*, SMO'07, pages 69–74, Stevens Point, Wisconsin, USA, 2007. World Scientific and Engineering Academy and Society (WSEAS).

[JMMS] M. Jalali, N. Mustapha, A. Mamat, and N.B. Sulaiman. A new classification model for online predicting users' future movements. In *Information Technology, 2008. ITSim 2008. International Symposium on*, pages 1–7, Aug.

[JMSM10] Mehrdad Jalali, Norwati Mustapha, Md. Nasir Sulaiman, and Ali Mamat. Webpum: A web-based recommendation system to predict user future movements. *Expert Systems with Applications*, 37(9):6201 – 6212, 2010.

[kOAa] kOA. Basic terms in advertising. http://www.knowonlineadvertising.com/facts-about-online-advertising/basic-terms-in-advertising/. Accessed last time on 9 Fev 2014.

[kOAb] kOA. Different products in online adv. http://www.knowonlineadvertising.com/facts-about-online-advertising/different-products-in-online-adv/. Accessed last time on 9 Fev 2014.

[kOAc] kOA. Standard ad sizes. http://www.knowonlineadvertising.com/facts-about-online-advertising/common-sizes-of-ads/. Accessed last time on 9 Fev 2014.

[LWT05] Chong Liu, Kui Wu, and Min Tsao. Energy efficient information collection with the arima model in wireless sensor networks. In *Global Telecommunications Conference, 2005. GLOBECOM '05. IEEE*, volume 5, pages 5 pp.–2474, Dec 2005.

[Mad12] T. Soni Madhulatha. An overview on clustering methods. *CoRR*, abs/1205.1117, 2012.

[Mak93] Spyros Makridakis. Accuracy measures: theoretical and practical concerns. *International Journal of Forecasting*, 9(4):527–529, December 1993.

[McN01] James McNames. A fast nearest-neighbor algorithm based on a principal axis search tree. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(9):964–976, Sep 2001.

[MHS+13] H. Brendan McMahan, Gary Holt, D. Sculley, Michael Young, Dietmar Ebner, Julian Grady, Lan Nie, Todd Phillips, Eugene Davydov, Daniel Golovin, Sharat Chikkerur, Dan Liu, Martin Wattenberg, Arnar Mar Hrafnkelsson, Tom Boulos, and Jeremy Kubica. Ad click prediction: A view from the trenches. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 1222–1230, New York, NY, USA, 2013. ACM.

[Nig99] Kamal Nigam. Using maximum entropy for text classification. In *In IJCAI-99 Workshop on Machine Learning for Information Filtering*, pages 61–67, 1999.

[NKM01] Alexandros Nanopoulos, Dimitris Katsaros, and Yannis Manolopoulos. Effective prediction of web-user accesses: A data mining approach, 2001.

[NMJ+13] Abhirup Nath, Shibnath Mukherjee, Prateek Jain, Navin Goyal, and Srivatsan Laxman. Ad impression forecasting for sponsored search. In *Proceedings of the 22nd international conference on World Wide Web*, pages 943–952. International World Wide Web Conferences Steering Committee, 2013.

[Per] DoubleClick Performics. Q1 2007 search trend report. http://www.swissmediatool.ch/_files/researchDB/228.pdf. Accessed last time on 10 Fev 2014.

[Pri13a] I.A.B. PricewaterhouseCoopers. Iab internet advertising revenue report, 2012 full year results. http://www.iab.net/media/file/IAB_Internet_Advertising_Revenue_Report_FY_2012_rev.pdf, April 2013. Accessed last time on 9 Fev 2014.

[Pri13b] I.A.B. PricewaterhouseCoopers. Q3 2013 internet advertising revenues climb to landmark high of nearly $10.7 billion, marking 15% year-over-year growth. http://www.iab.net/about_the_iab/recent_press_releases/press_release_archive/press_release/pr-122313, December 2013. Accessed last time on 9 Fev 2014.

[Sab07] M Sabry. Comparison between regression and arima models in forecasting traffic volume. *Australian Journal of Basic and Applied Sciences*, 1(2):126, 2007.

[Sas07] Yutaka Sasaki. The truth of the f-measure. *Teaching, Tutorial materials, Version: 26th October*, 2007.

[Spr91] RobertF. Sproull. Refinements to nearest-neighbor searching ink-dimensional trees. *Algorithmica*, 6(1-6):579–589, 1991.

[TOY+13] Yukihiro Tagami, Shingo Ono, Koji Yamamoto, Koji Tsukamoto, and Akira Tajima. Ctr prediction for contextual advertising: Learning-to-rank approach. In *Proceedings of the Seventh International Workshop on Data Mining for Online Advertising*, ADKDD '13, pages 4:1–4:8, New York, NY, USA, 2013. ACM.

[UP12] Sachin Pardeshi Ujwala Patil. A survey on user future request prediction: Web usage mining. *International Journal of Emerging Technology and Advanced Engineering (ISSN 2250-2459, Volume 2, Issue 3, March 2012)*, 2, 2012.

[WF05] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005.

[WK06] C. Walgampaya and M. Kantardzic. Selection of distributed sensors for multiple time series prediction. In *Neural Networks, 2006. IJCNN '06. International Joint Conference on*, pages 3152–3158, 2006.

[YLGX10] Ruixi Yuan, Zhu Li, Xiaohong Guan, and Li Xu. An svm-based machine learning method for accurate internet traffic classification. *Information Systems Frontiers*, 12(2):149–156, 2010.

[YWZ13] Shuai Yuan, Jun Wang, and Xiaoxue Zhao. Real-time bidding for online advertising: Measurement and analysis. In *Proceedings of the Seventh International Workshop on Data Mining for Online Advertising*, ADKDD '13, pages 3:1–3:8, New York, NY, USA, 2013. ACM.

[ZLX09] Yong Zhou, Youwen Li, and Shixiong Xia. An improved knn text classification algorithm based on clustering. *JCP*, 4(3):230–237, 2009.

Appendix A

Case 1

A.1 Baseline - 4h

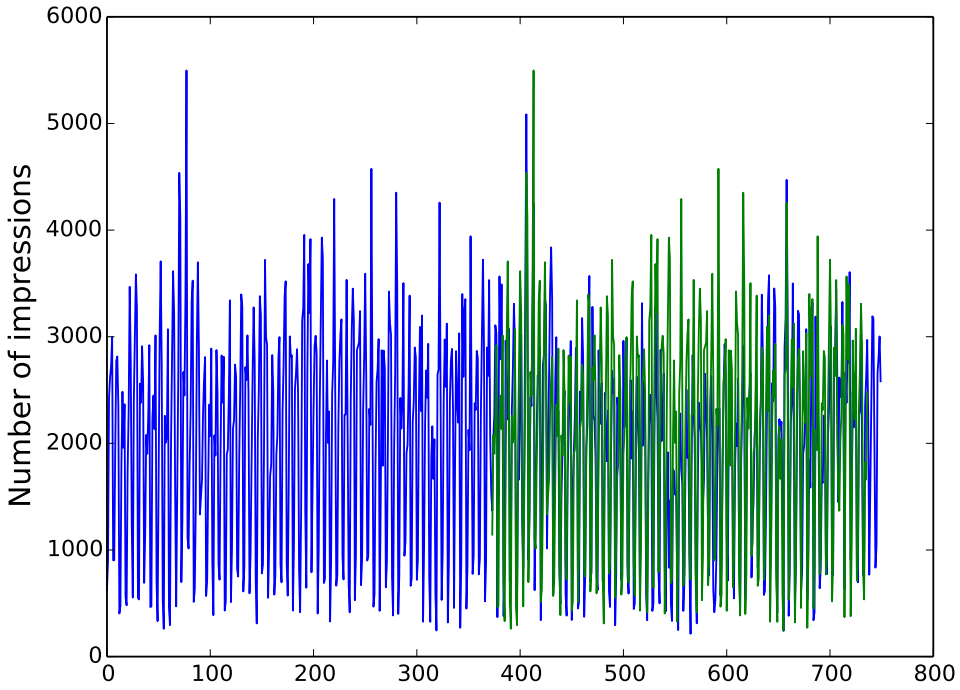

Figure A.1: Baseline - Impressions forecast from 2013-11-26 04:00:00 to 2014-01-25 16:00:00

Case 1

σ (Real Data)	RMSE	MASE
946.25	690.87	0.5227

Table A.1: Baseline - Error for Impressions forecast from 2013-11-26 04:00:00 to 2014-01-25 16:00:00

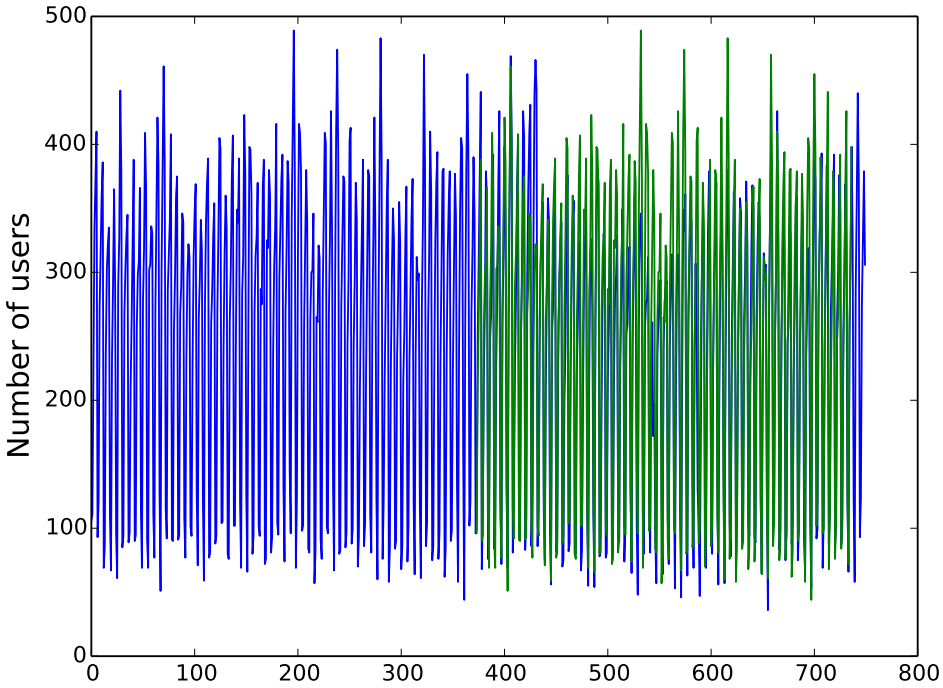

Figure A.2: Baseline - Uniques forecast from 2013-11-26 04:00:00 to 2014-01-25 16:00:00

σ (Real Data)	RMSE	MASE
113.42	50.02	0.3337

Table A.2: Baseline - Error for Uniques forecast from 2013-11-26 04:00:00 to 2014-01-25 16:00:00

Case 1

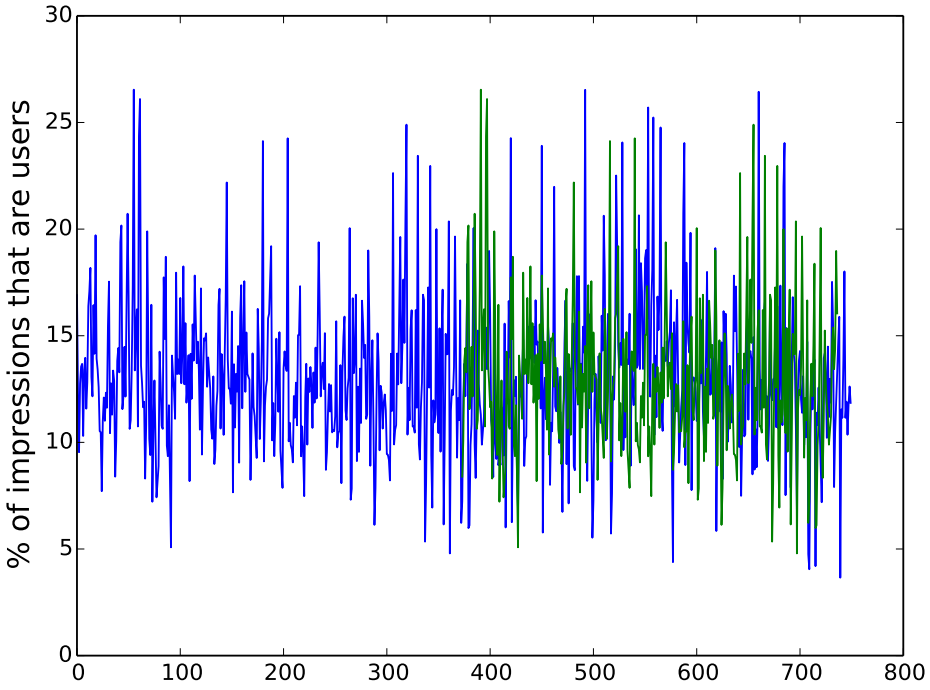

Figure A.3: Baseline - Uniques Percentage forecast from 2013-11-26 04:00:00 to 2014-01-25 16:00:00

σ (Real Data)	RMSE	MASE
3.69	4.64	1.0046

Table A.3: Baseline - Error for Uniques Percentage forecast from 2013-11-26 04:00:00 to 2014-01-25 16:00:00

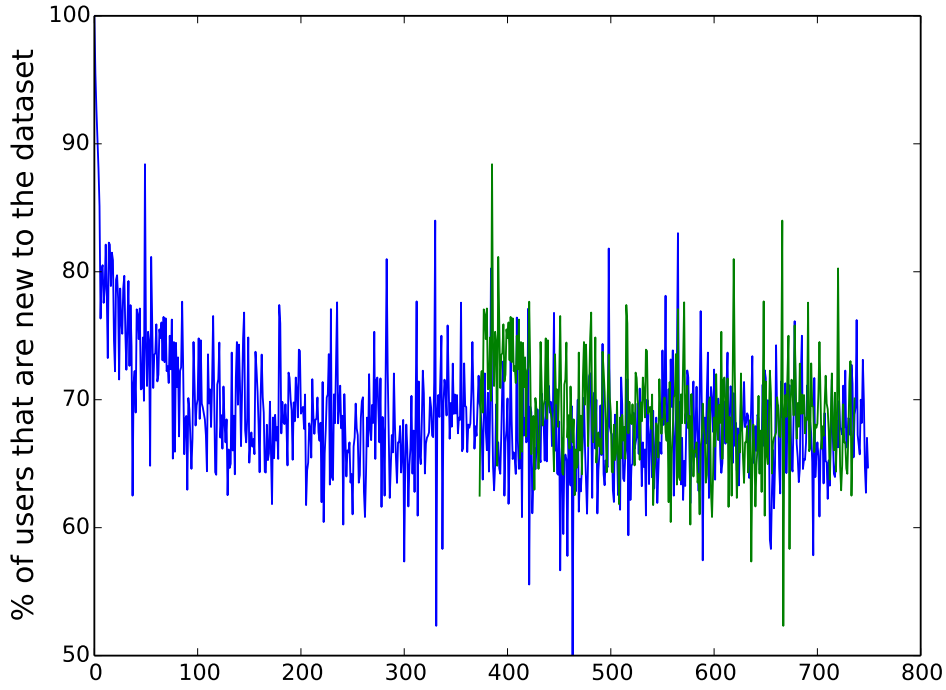

Figure A.4: Baseline - New uniques forecast from 2013-11-26 04:00:00 to 2014-01-25 16:00:00

σ (Real Data)	RMSE	MASE
3.83	5.88	1.1278

Table A.4: Baseline - Error for New Uniques forecast from 2013-11-26 04:00:00 to 2014-01-25 16:00:00

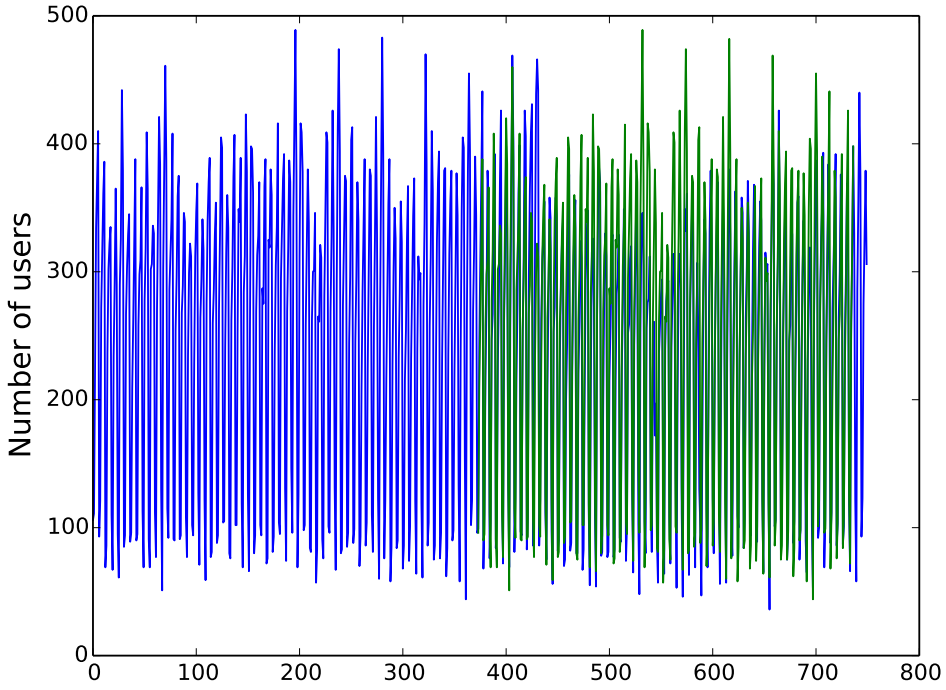

Figure A.5: Baseline - Uniques calculated using percentages forecast from 2013-11-26 04:00:00 to 2014-01-25 16:00:00

σ (Real Data)	RMSE	MASE
113.42	49.98	0.3333

Table A.5: Baseline - Error for Uniques calculated using percentages forecast from 2013-11-26 04:00:00 to 2014-01-25 16:00:00

A.2 Arima Allow Drift True - 4h

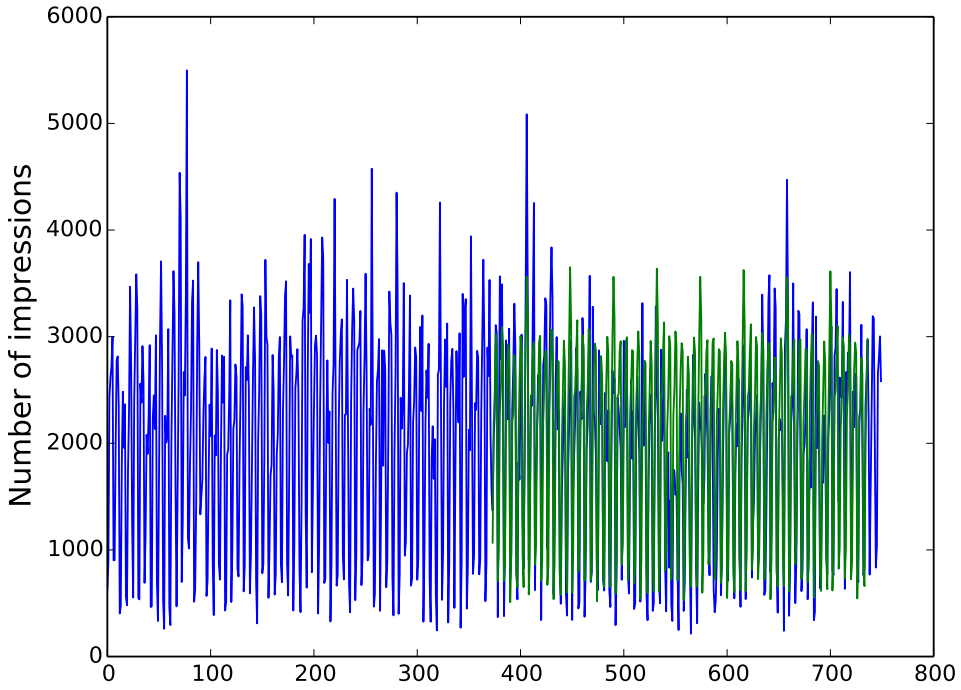

Figure A.6: Arima Allow Drift True - Impressions forecast from 2013-11-26 04:00:00 to 2014-01-25 16:00:00

σ (Real Data)	RMSE	MASE
946.25	562.8	0.4397

Table A.6: Arima Allow Drift True - Error for Impressions forecast from 2013-11-26 04:00:00 to 2014-01-25 16:00:00

Case 1

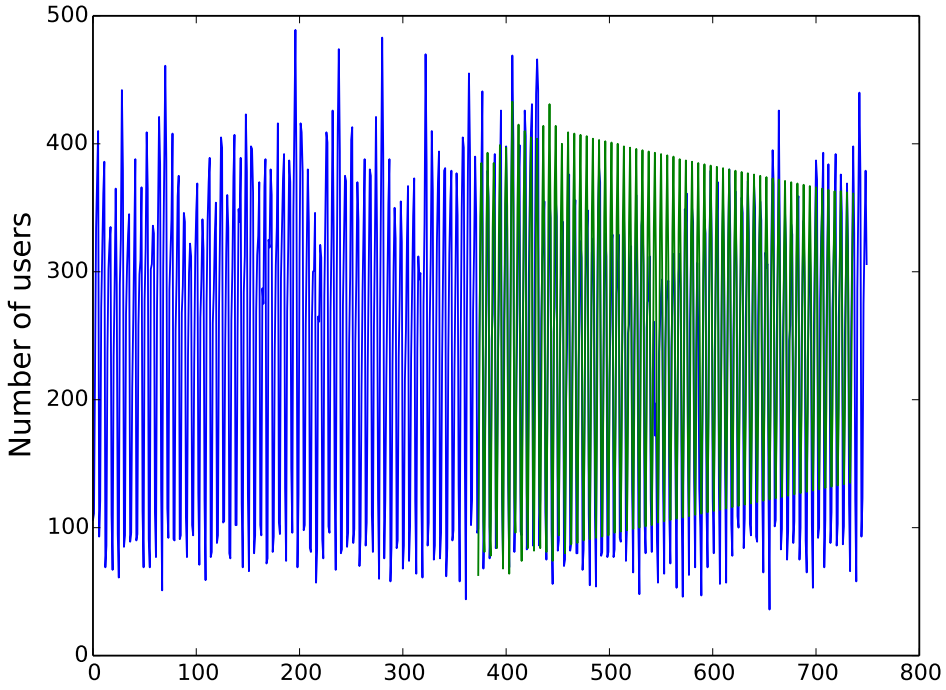

Figure A.7: Arima Allow Drift True - Uniques forecast from 2013-11-26 04:00:00 to 2014-01-25 16:00:00

σ (Real Data)	RMSE	MASE
113.42	53.77	0.4115

Table A.7: Arima Allow Drift True - Error for Uniques forecast from 2013-11-26 04:00:00 to 2014-01-25 16:00:00

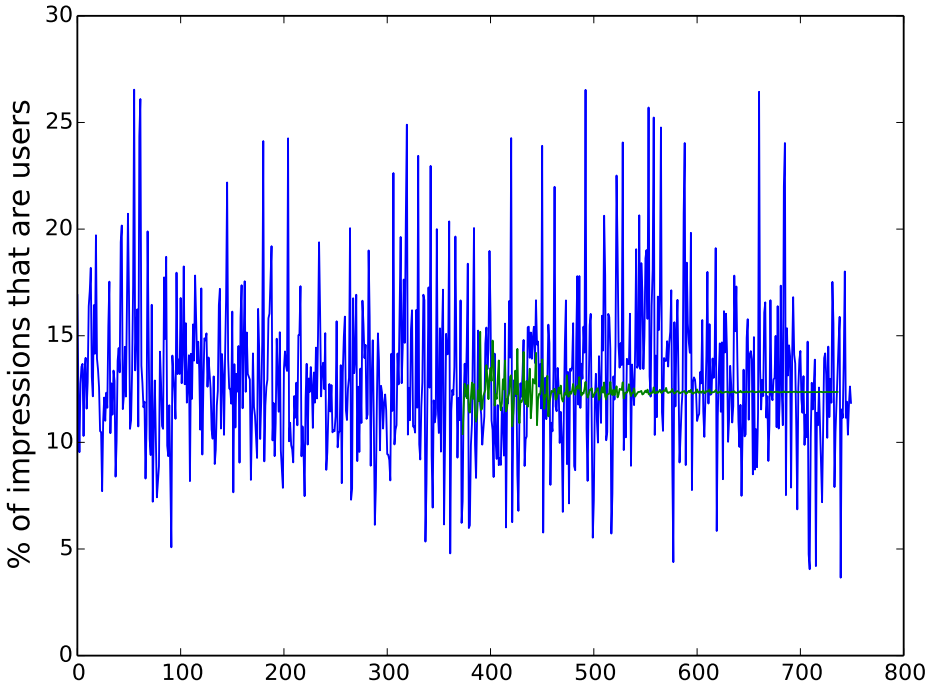

Figure A.8: Arima Allow Drift True - Uniques Percentage forecast from 2013-11-26 04:00:00 to 2014-01-25 16:00:00

σ (Real Data)	RMSE	MASE
3.69	3.7	0.7885

Table A.8: Arima Allow Drift True - Error for Uniques Percentage forecast from 2013-11-26 04:00:00 to 2014-01-25 16:00:00

Case 1

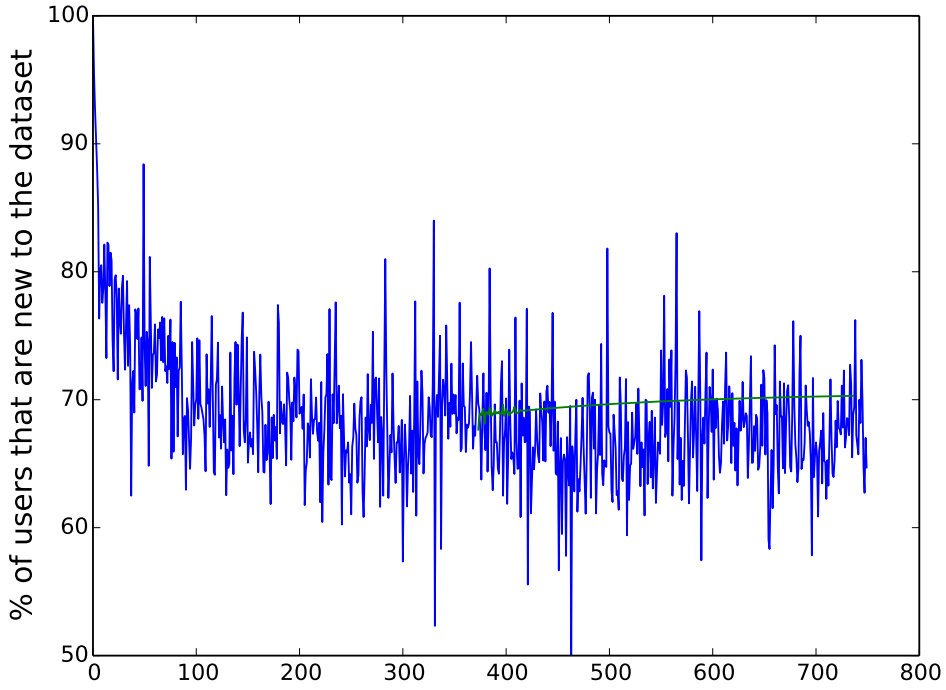

Figure A.9: Arima Allow Drift True - New uniques forecast from 2013-11-26 04:00:00 to 2014-01-25 16:00:00

σ (Real Data)	RMSE	MASE
3.83	4.64	0.9139

Table A.9: Arima Allow Drift True - Error for New Uniques forecast from 2013-11-26 04:00:00 to 2014-01-25 16:00:00

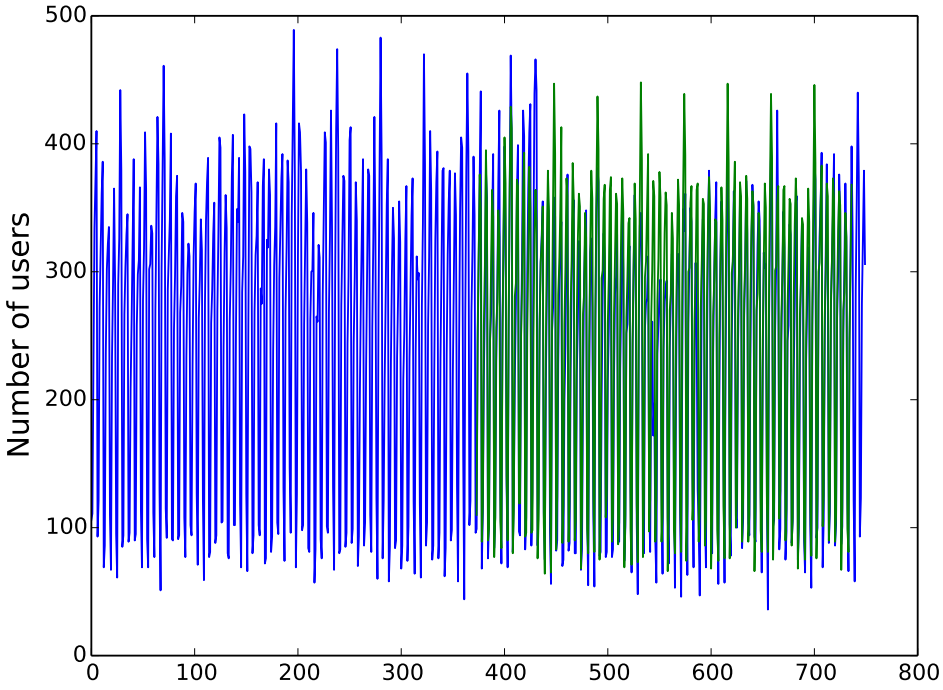

Figure A.10: Arima Allow Drift True - Uniques calculated using percentages forecast from 2013-11-26 04:00:00 to 2014-01-25 16:00:00

σ (Real Data)	RMSE	MASE
113.42	41.87	0.2878

Table A.10: Arima Allow Drift True - Error for Uniques calculated using percentages forecast from 2013-11-26 04:00:00 to 2014-01-25 16:00:00

A.3 Arima Allow Drift False - 4h

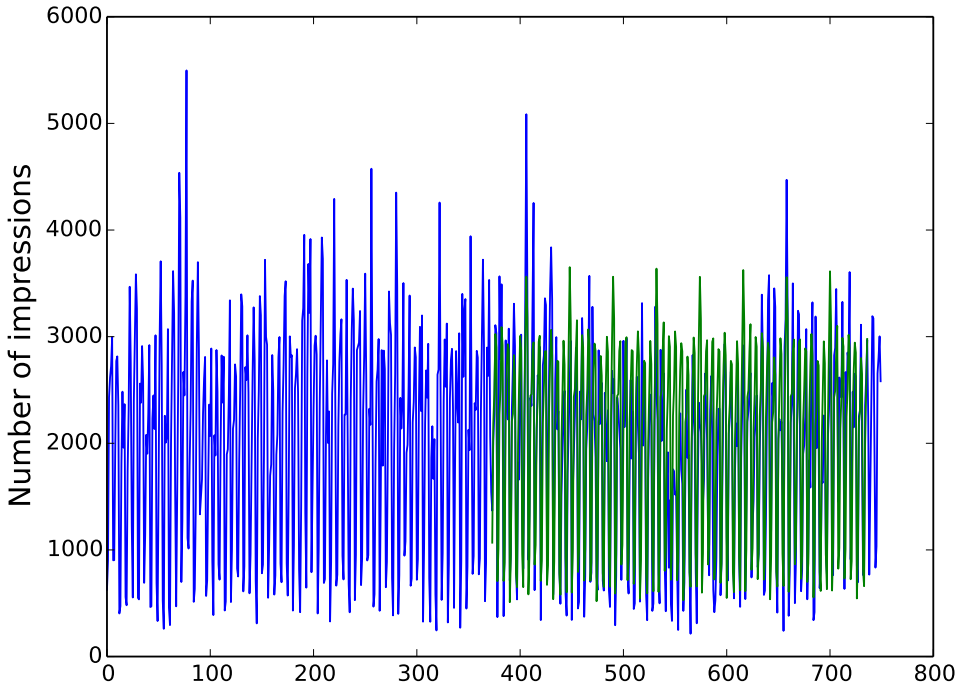

Figure A.11: Arima Allow Drift False - Impressions forecast from 2013-11-26 04:00:00 to 2014-01-25 16:00:00

σ (Real Data)	RMSE	MASE
946.25	562.8	0.4397

Table A.11: Arima Allow Drift False - Error for Impressions forecast from 2013-11-26 04:00:00 to 2014-01-25 16:00:00

Case 1

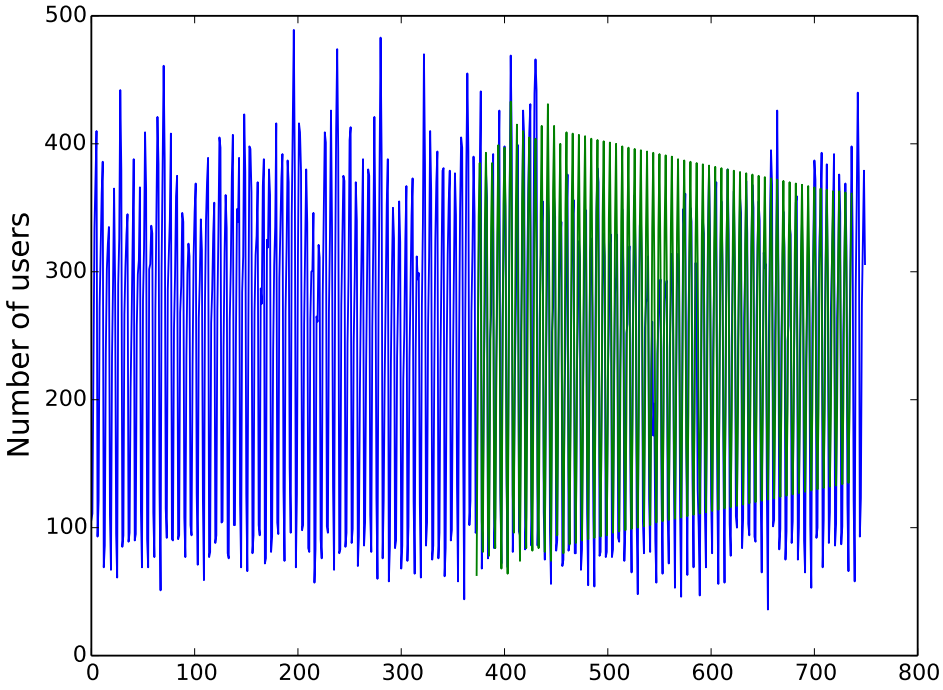

Figure A.12: Arima Allow Drift False - Uniques forecast from 2013-11-26 04:00:00 to 2014-01-25 16:00:00

σ (Real Data)	RMSE	MASE
113.42	53.77	0.4115

Table A.12: Arima Allow Drift False - Error for Uniques forecast from 2013-11-26 04:00:00 to 2014-01-25 16:00:00

Case 1

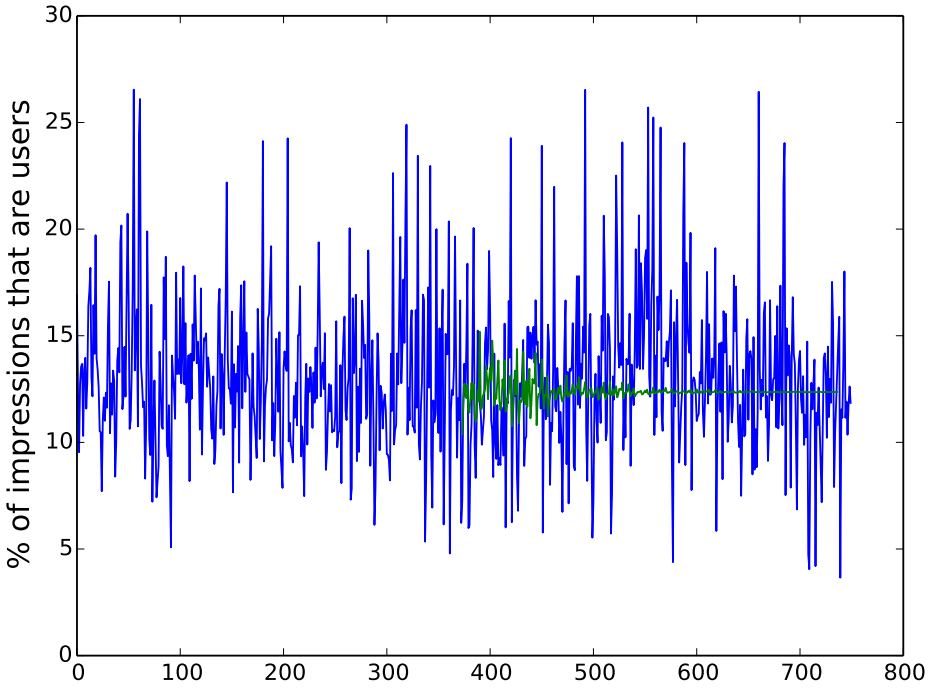

Figure A.13: Arima Allow Drift False - Uniques Percentage forecast from 2013-11-26 04:00:00 to 2014-01-25 16:00:00

σ (Real Data)	RMSE	MASE
3.69	3.7	0.7885

Table A.13: Arima Allow Drift False - Error for Uniques Percentage forecast from 2013-11-26 04:00:00 to 2014-01-25 16:00:00

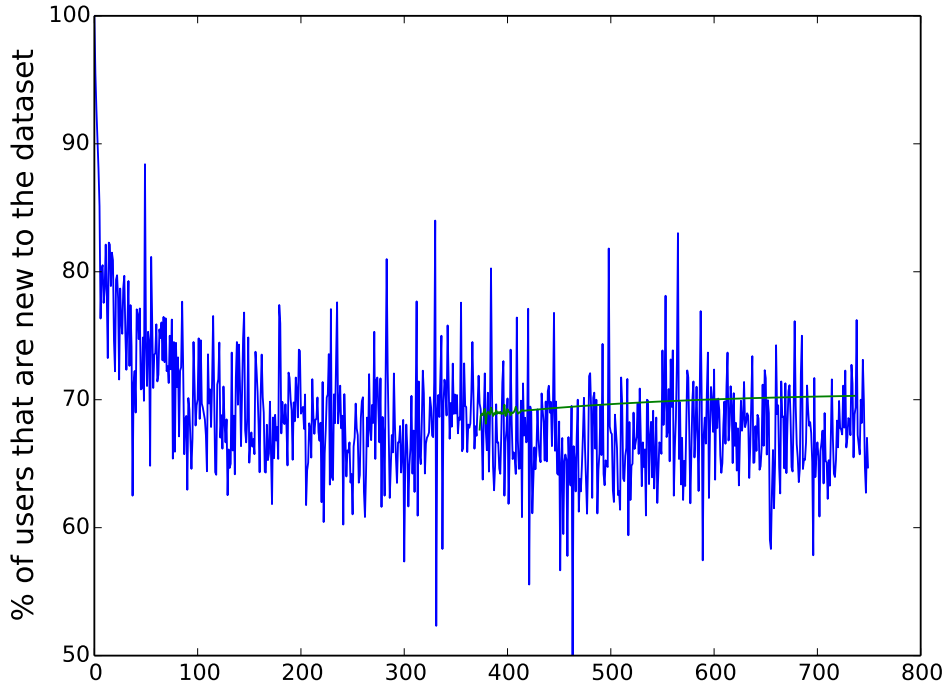

Figure A.14: Arima Allow Drift False - New uniques forecast from 2013-11-26 04:00:00 to 2014-01-25 16:00:00

σ (Real Data)	RMSE	MASE
3.83	4.64	0.9139

Table A.14: Arima Allow Drift False - Error for New Uniques forecast from 2013-11-26 04:00:00 to 2014-01-25 16:00:00

Case 1

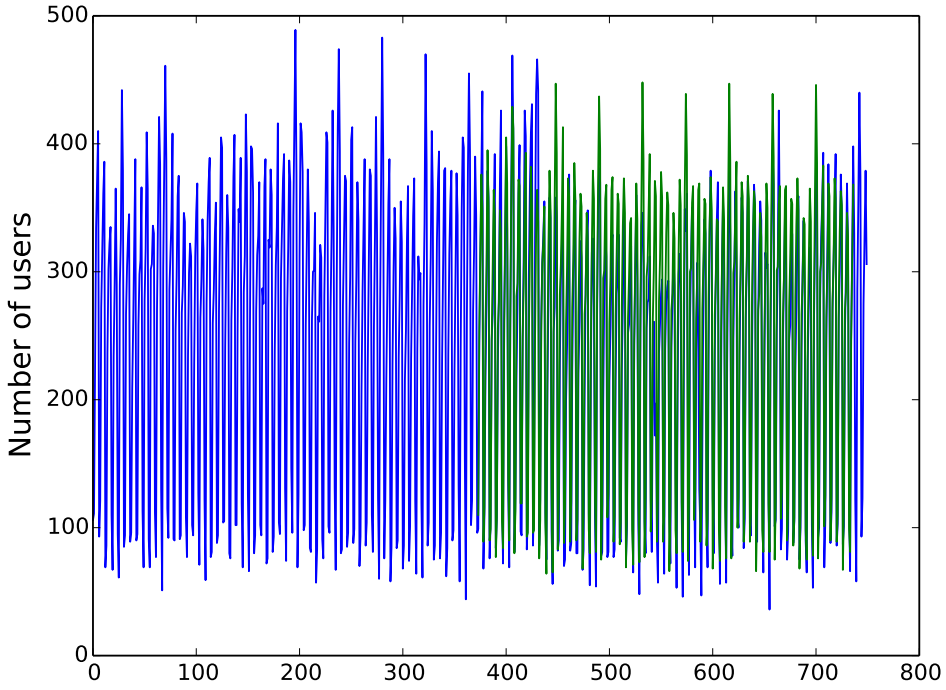

Figure A.15: Arima Allow Drift False - Uniques calculated using percentages forecast from 2013-11-26 04:00:00 to 2014-01-25 16:00:00

σ (Real Data)	RMSE	MASE
113.42	41.87	0.2878

Table A.15: Arima Allow Drift False - Error for Uniques calculated using percentages forecast from 2013-11-26 04:00:00 to 2014-01-25 16:00:00

A.4 Baseline - 6h

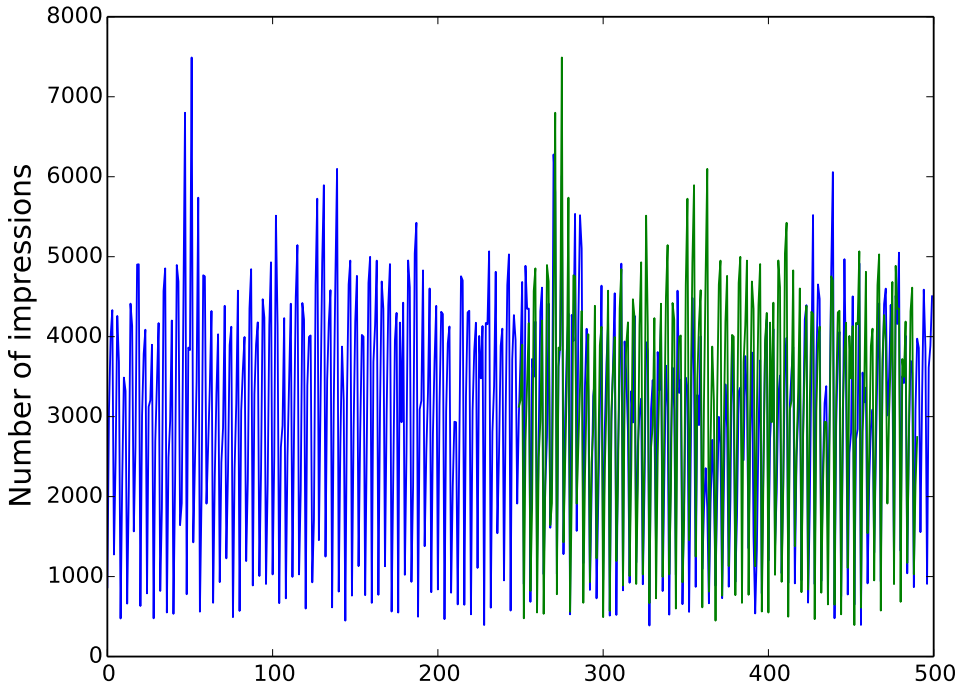

Figure A.16: Baseline - Impressions forecast from 2013-11-26 06:00:00 to 2014-01-25 12:00:00

σ (Real Data)	RMSE	MASE
1358.24	945.7	0.3659

Table A.16: Baseline - Error for Impressions forecast from 2013-11-26 06:00:00 to 2014-01-25 12:00:00

Case 1

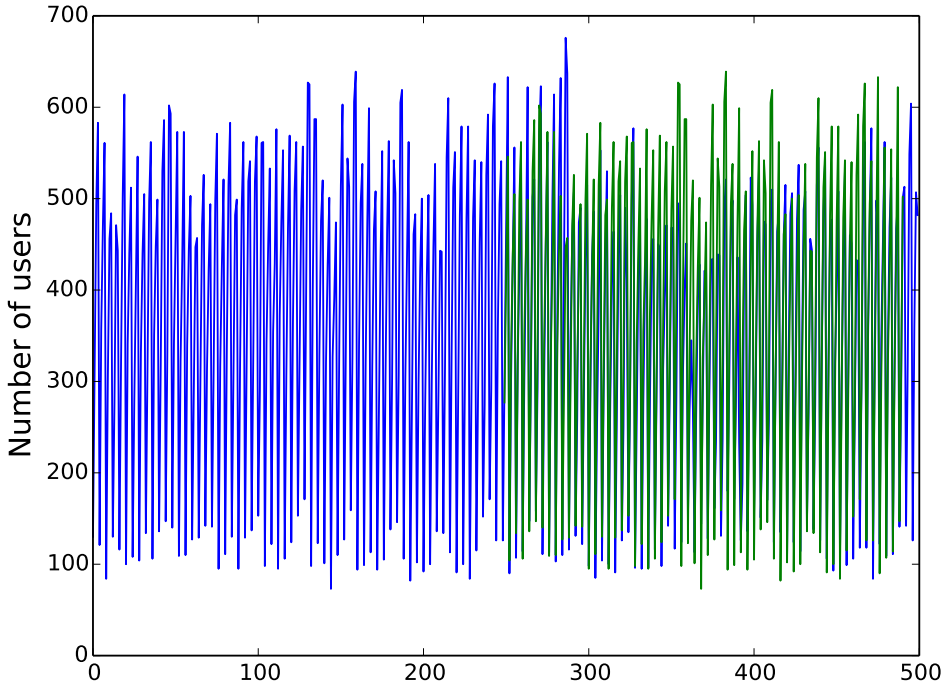

Figure A.17: Baseline - Uniques forecast from 2013-11-26 06:00:00 to 2014-01-25 12:00:00

σ (Real Data)	RMSE	MASE
158.41	68.65	0.2203

Table A.17: Baseline - Error for Uniques forecast from 2013-11-26 06:00:00 to 2014-01-25 12:00:00

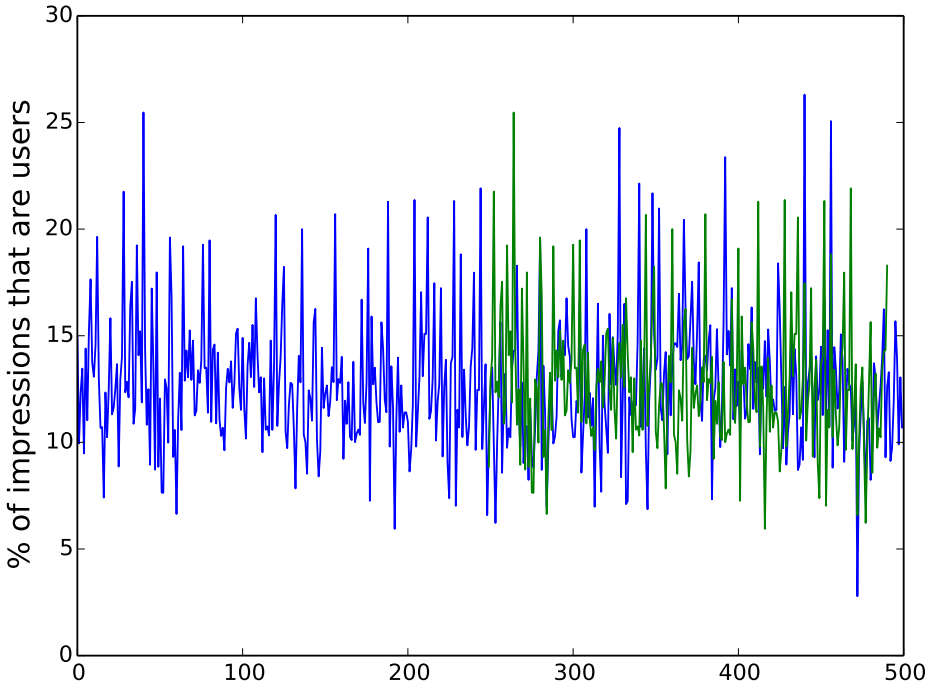

Figure A.18: Baseline - Uniques Percentage forecast from 2013-11-26 06:00:00 to 2014-01-25 12:00:00

σ (Real Data)	RMSE	MASE
3.15	4.1	0.8604

Table A.18: Baseline - Error for Uniques Percentage forecast from 2013-11-26 06:00:00 to 2014-01-25 12:00:00

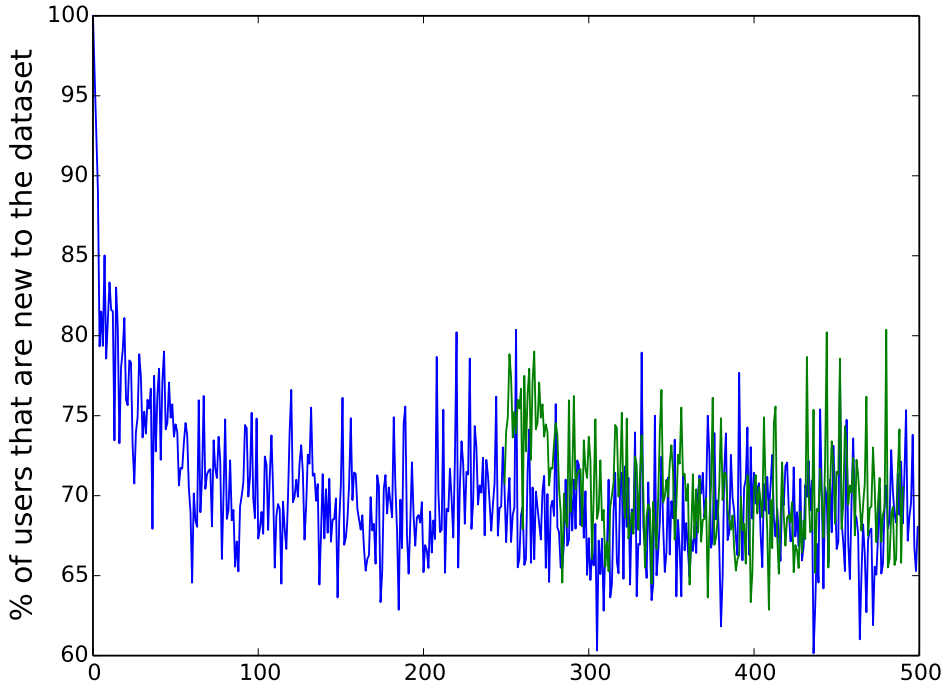

Figure A.19: Baseline - New uniques forecast from 2013-11-26 06:00:00 to 2014-01-25 12:00:00

σ (Real Data)	RMSE	MASE
3.15	4.87	1.1985

Table A.19: Baseline - Error for New Uniques forecast from 2013-11-26 06:00:00 to 2014-01-25 12:00:00

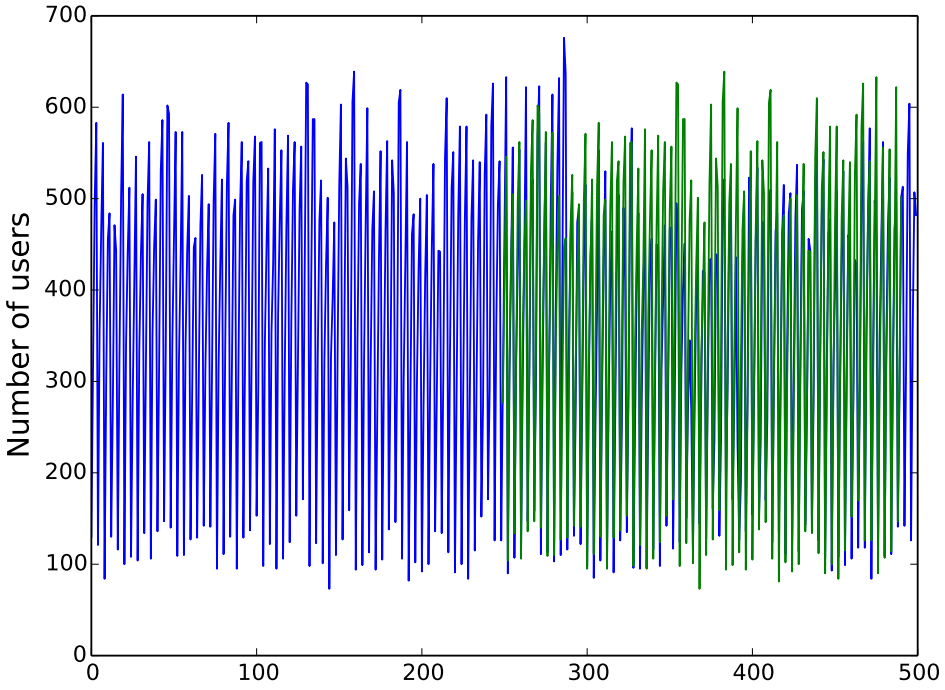

Figure A.20: Baseline - Uniques calculated using percentages forecast from 2013-11-26 06:00:00 to 2014-01-25 12:00:00

σ (Real Data)	RMSE	MASE
158.41	68.65	0.2203

Table A.20: Baseline - Error for Uniques calculated using percentages forecast from 2013-11-26 06:00:00 to 2014-01-25 12:00:00

A.5 Arima Allow Drift True - 6h

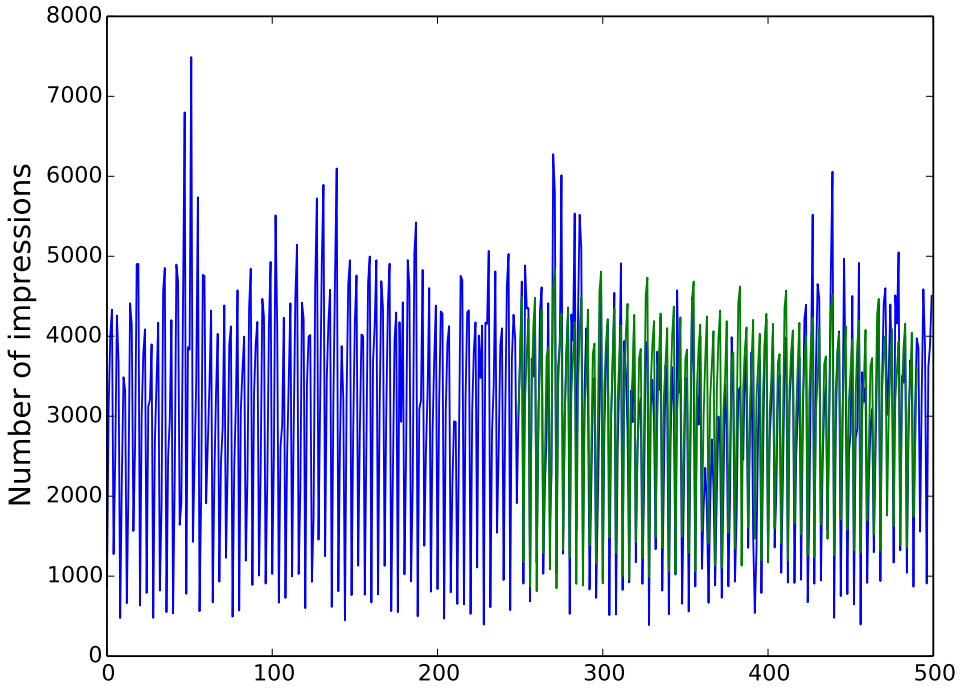

Figure A.21: Arima Allow Drift True - Impressions forecast from 2013-11-26 06:00:00 to 2014-01-25 12:00:00

σ (Real Data)	RMSE	MASE
1358.24	742.57	0.3003

Table A.21: Arima Allow Drift True - Error for Impressions forecast from 2013-11-26 06:00:00 to 2014-01-25 12:00:00

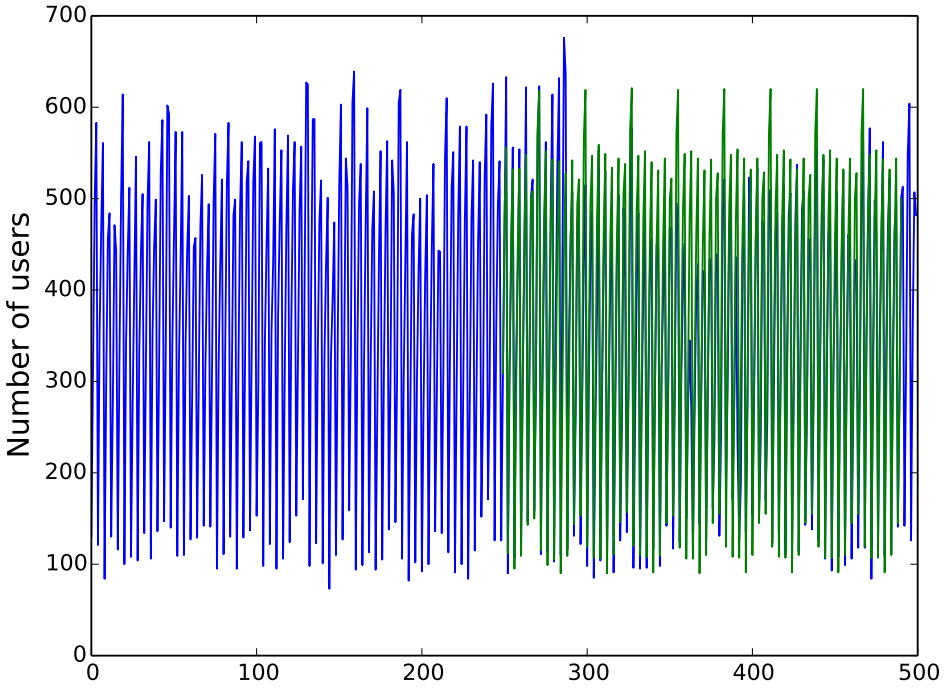

Figure A.22: Arima Allow Drift True - Uniques forecast from 2013-11-26 06:00:00 to 2014-01-25 12:00:00

σ (Real Data)	RMSE	MASE
158.41	57.26	0.1772

Table A.22: Arima Allow Drift True - Error for Uniques forecast from 2013-11-26 06:00:00 to 2014-01-25 12:00:00

Case 1

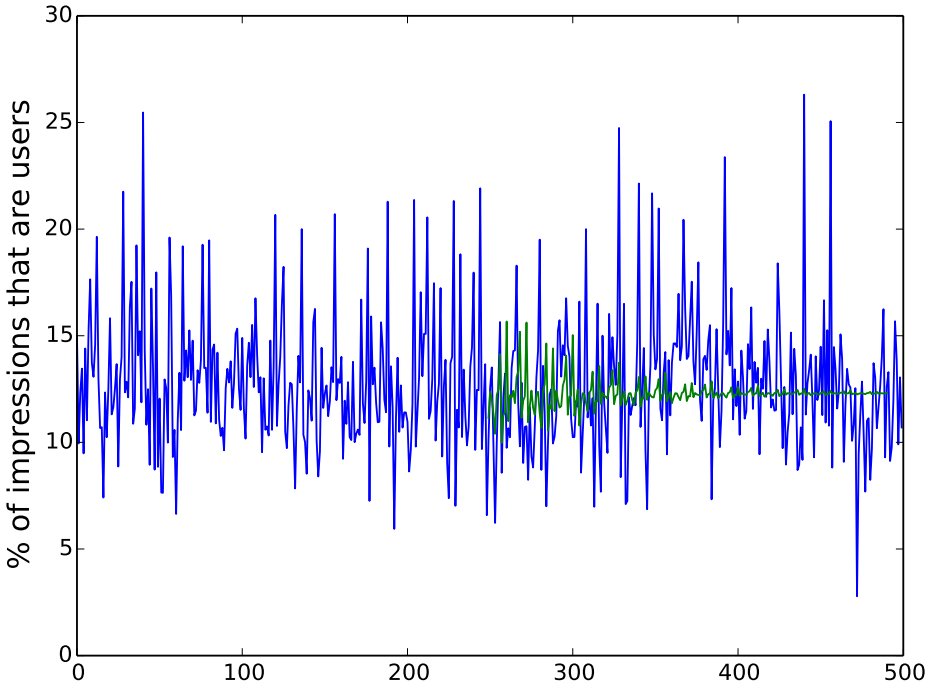

Figure A.23: Arima Allow Drift True - Uniques Percentage forecast from 2013-11-26 06:00:00 to 2014-01-25 12:00:00

σ (Real Data)	RMSE	MASE
3.15	3.2	0.6554

Table A.23: Arima Allow Drift True - Error for Uniques Percentage forecast from 2013-11-26 06:00:00 to 2014-01-25 12:00:00

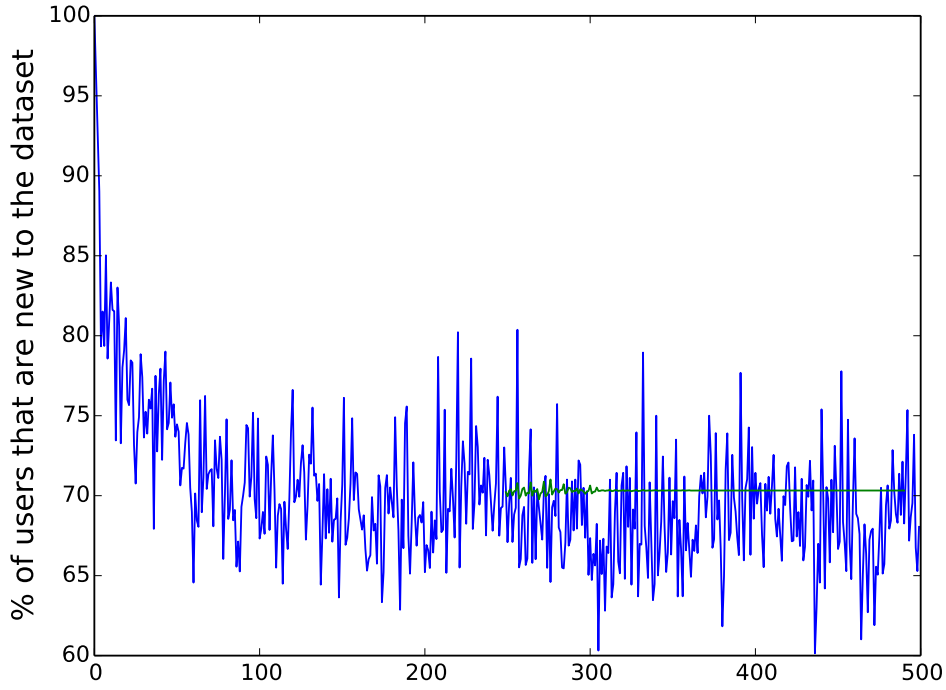

Figure A.24: Arima Allow Drift True - New uniques forecast from 2013-11-26 06:00:00 to 2014-01-25 12:00:00

σ (Real Data)	RMSE	MASE
3.15	3.6	0.9087

Table A.24: Arima Allow Drift True - Error for New Uniques forecast from 2013-11-26 06:00:00 to 2014-01-25 12:00:00

Case 1

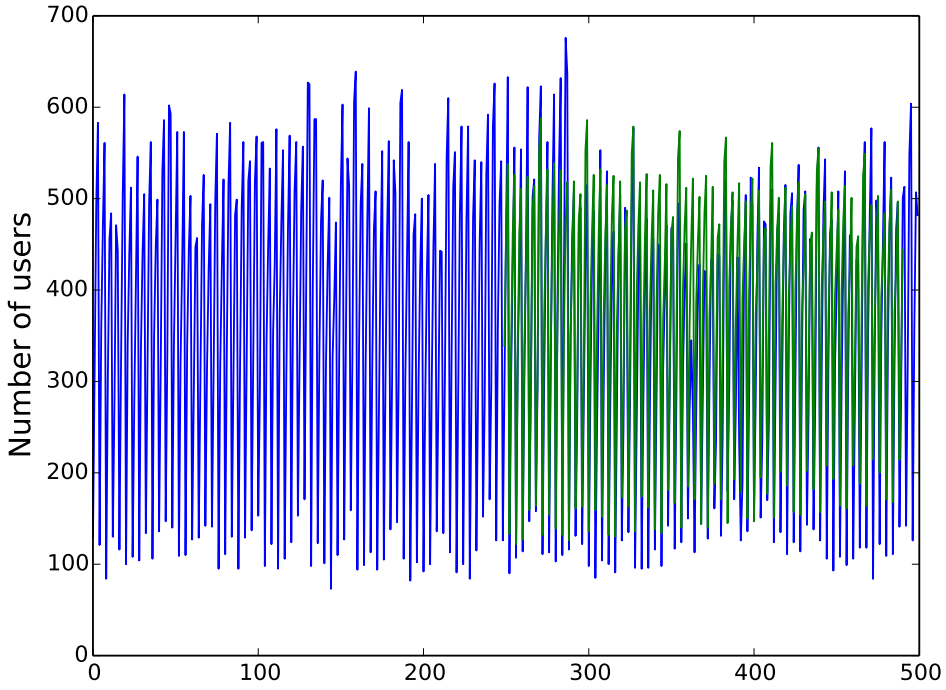

Figure A.25: Arima Allow Drift True - Uniques calculated using percentages forecast from 2013-11-26 06:00:00 to 2014-01-25 12:00:00

σ (Real Data)	RMSE	MASE
158.41	63.41	0.2224

Table A.25: Arima Allow Drift True - Error for Uniques calculated using percentages forecast from 2013-11-26 06:00:00 to 2014-01-25 12:00:00

A.6 Arima Allow Drift False - 6h

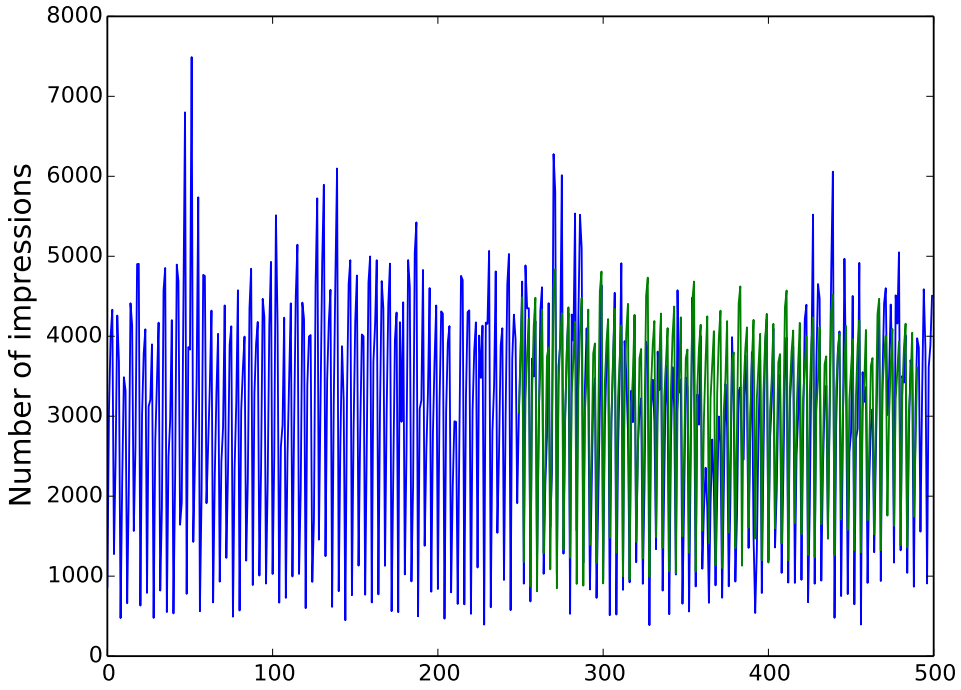

Figure A.26: Arima Allow Drift False - Impressions forecast from 2013-11-26 06:00:00 to 2014-01-25 12:00:00

σ (Real Data)	RMSE	MASE
1358.24	742.57	0.3003

Table A.26: Arima Allow Drift False - Error for Impressions forecast from 2013-11-26 06:00:00 to 2014-01-25 12:00:00

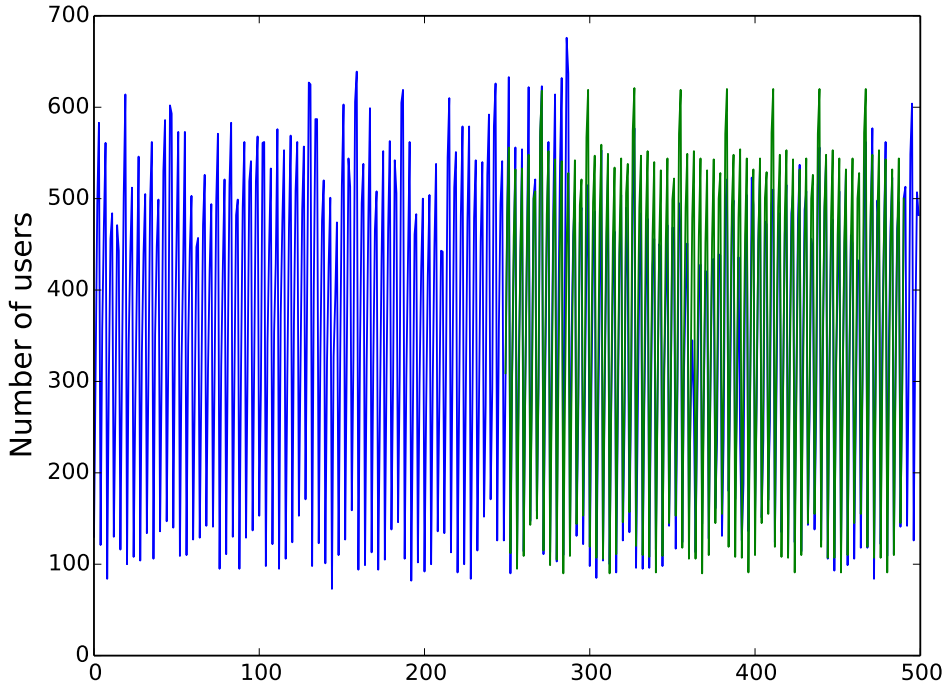

Figure A.27: Arima Allow Drift False - Uniques forecast from 2013-11-26 06:00:00 to 2014-01-25 12:00:00

σ (Real Data)	RMSE	MASE
158.41	57.26	0.1772

Table A.27: Arima Allow Drift False - Error for Uniques forecast from 2013-11-26 06:00:00 to 2014-01-25 12:00:00

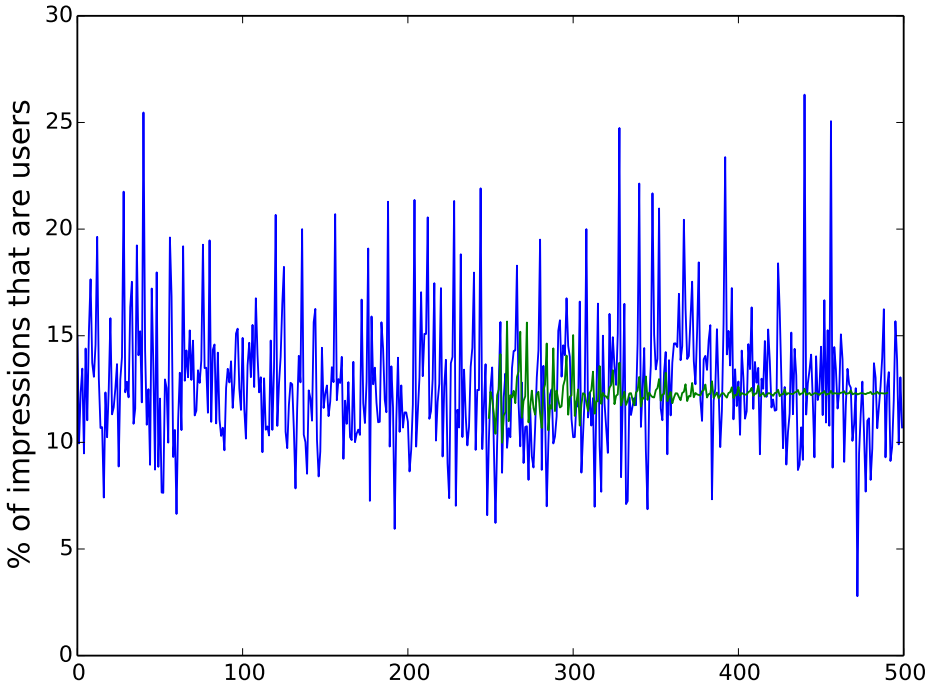

Figure A.28: Arima Allow Drift False - Uniques Percentage forecast from 2013-11-26 06:00:00 to 2014-01-25 12:00:00

σ (Real Data)	RMSE	MASE
3.15	3.2	0.6554

Table A.28: Arima Allow Drift False - Error for Uniques Percentage forecast from 2013-11-26 06:00:00 to 2014-01-25 12:00:00

Case 1

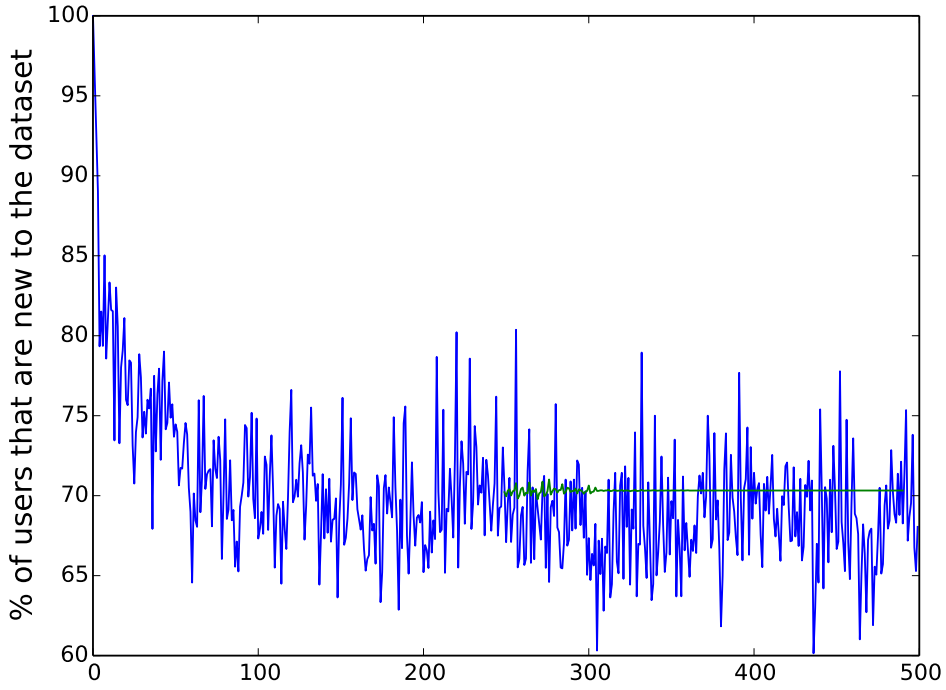

Figure A.29: Arima Allow Drift False - New uniques forecast from 2013-11-26 06:00:00 to 2014-01-25 12:00:00

σ (Real Data)	RMSE	MASE
3.15	3.6	0.9087

Table A.29: Arima Allow Drift False - Error for New Uniques forecast from 2013-11-26 06:00:00 to 2014-01-25 12:00:00

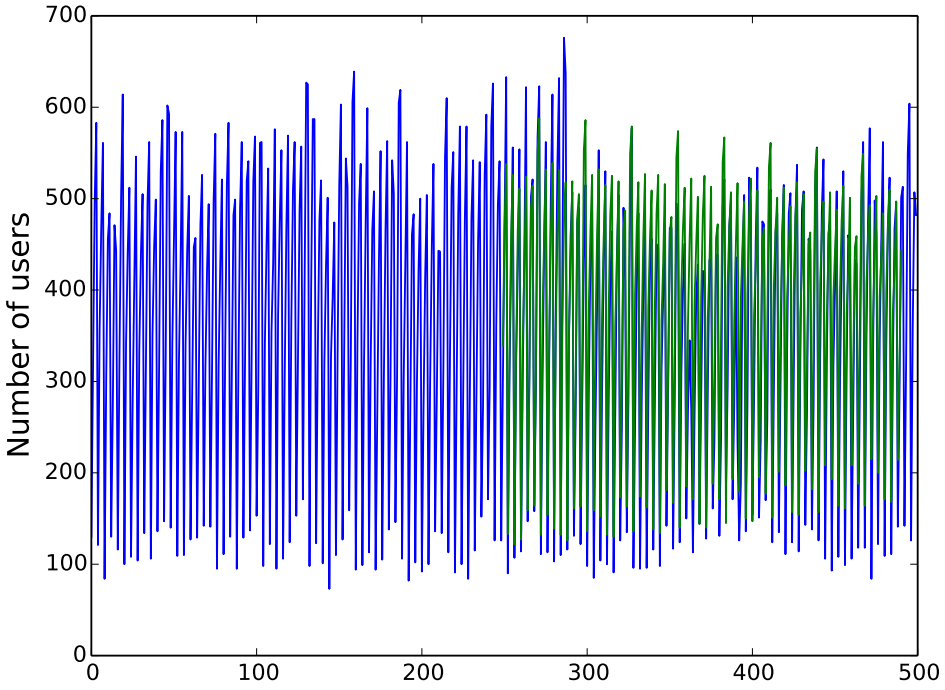

Figure A.30: Arima Allow Drift False - Uniques calculated using percentages forecast from 2013-11-26 06:00:00 to 2014-01-25 12:00:00

σ (Real Data)	RMSE	MASE
158.41	63.41	0.2224

Table A.30: Arima Allow Drift False - Error for Uniques calculated using percentages forecast from 2013-11-26 06:00:00 to 2014-01-25 12:00:00

A.7 Baseline - 8h

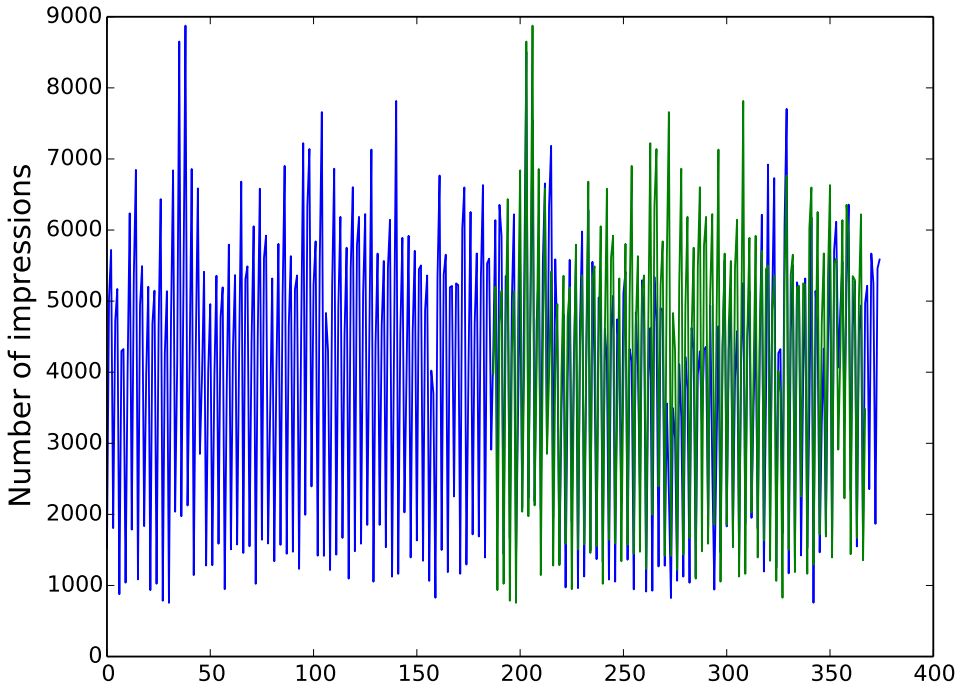

Figure A.31: Baseline - Impressions forecast from 2013-11-26 08:00:00 to 2014-01-25 08:00:00

σ (Real Data)	RMSE	MASE
1785.84	1150.44	0.2712

Table A.31: Baseline - Error for Impressions forecast from 2013-11-26 08:00:00 to 2014-01-25 08:00:00

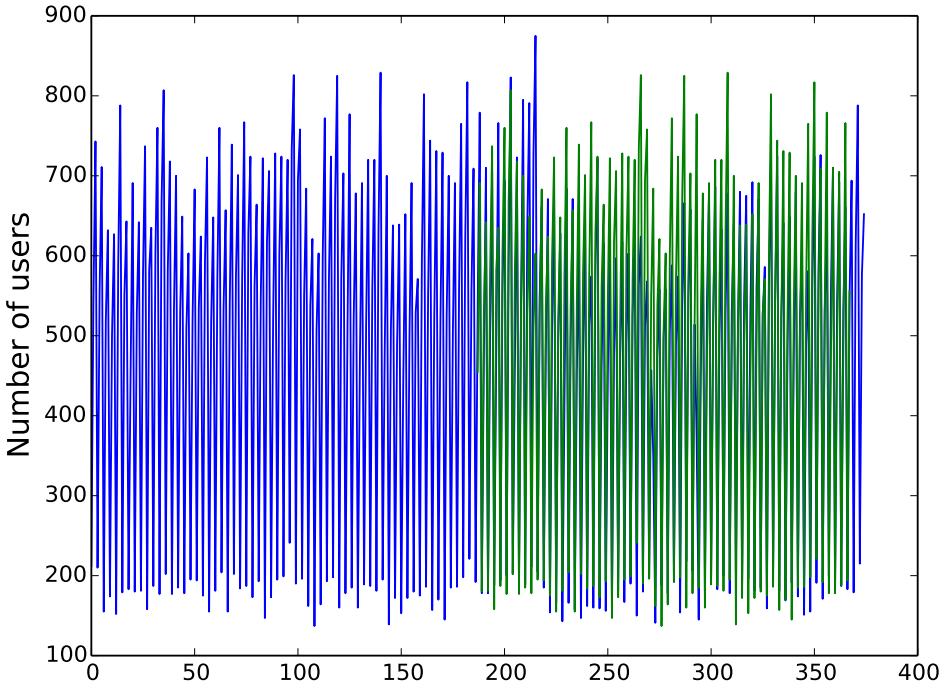

Figure A.32: Baseline - Uniques forecast from 2013-11-26 08:00:00 to 2014-01-25 08:00:00

σ (Real Data)	RMSE	MASE
209.5	84.57	0.1616

Table A.32: Baseline - Error for Uniques forecast from 2013-11-26 08:00:00 to 2014-01-25 08:00:00

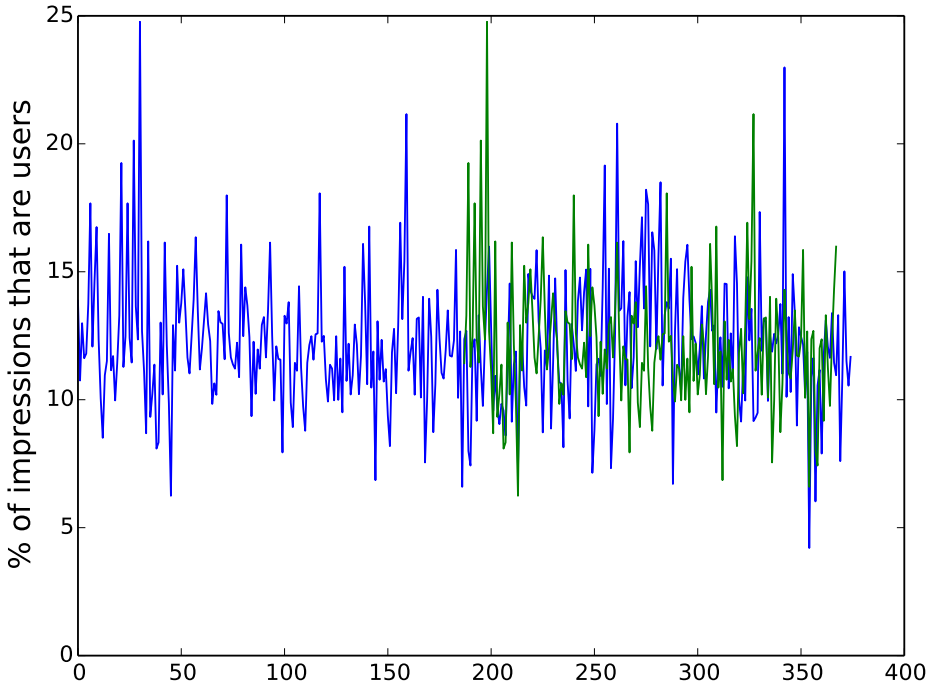

Figure A.33: Baseline - Uniques Percentage forecast from 2013-11-26 08:00:00 to 2014-01-25 08:00:00

σ (Real Data)	RMSE	MASE
2.66	3.53	0.9989

Table A.33: Baseline - Error for Uniques Percentage forecast from 2013-11-26 08:00:00 to 2014-01-25 08:00:00

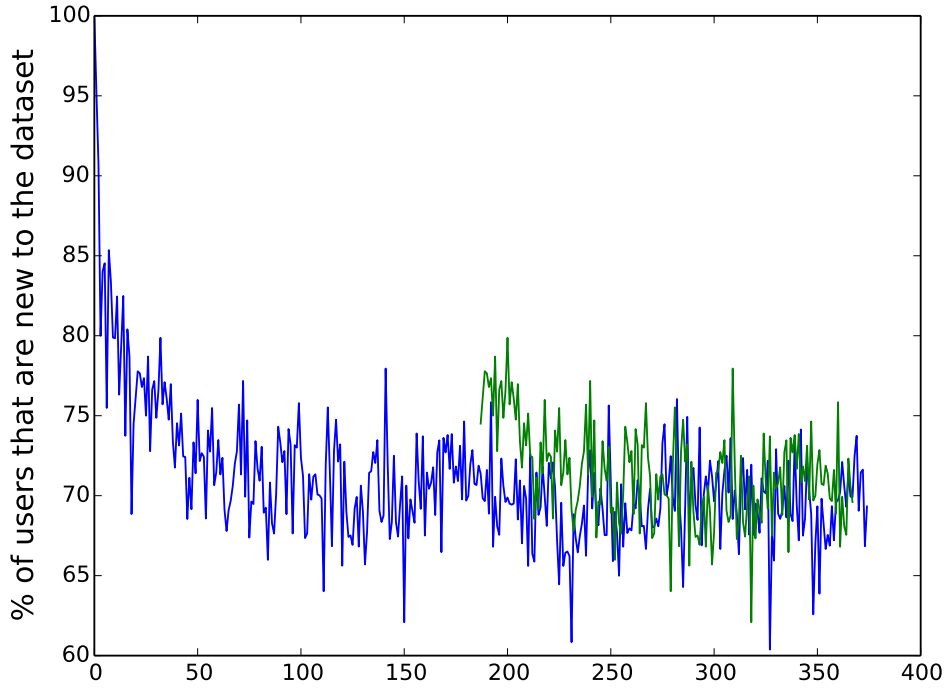

Figure A.34: Baseline - New uniques forecast from 2013-11-26 08:00:00 to 2014-01-25 08:00:00

σ (Real Data)	RMSE	MASE
2.47	4.5	1.2051

Table A.34: Baseline - Error for New Uniques forecast from 2013-11-26 08:00:00 to 2014-01-25 08:00:00

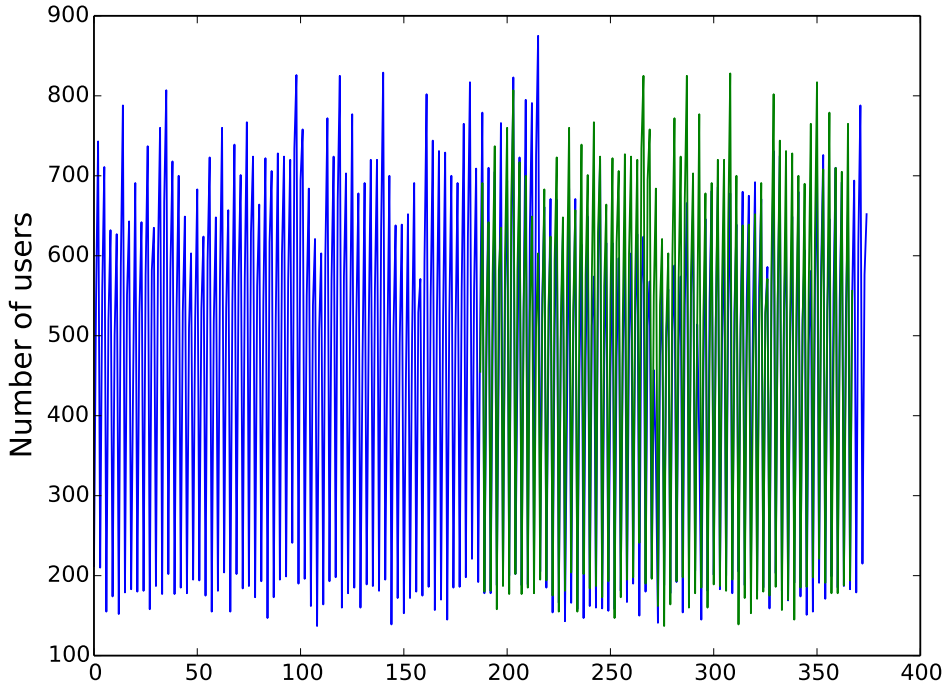

Figure A.35: Baseline - Uniques calculated using percentages forecast from 2013-11-26 08:00:00 to 2014-01-25 08:00:00

σ (Real Data)	RMSE	MASE
209.5	84.51	0.1615

Table A.35: Baseline - Error for Uniques calculated using percentages forecast from 2013-11-26 08:00:00 to 2014-01-25 08:00:00

A.8 Arima Allow Drift True - 8h

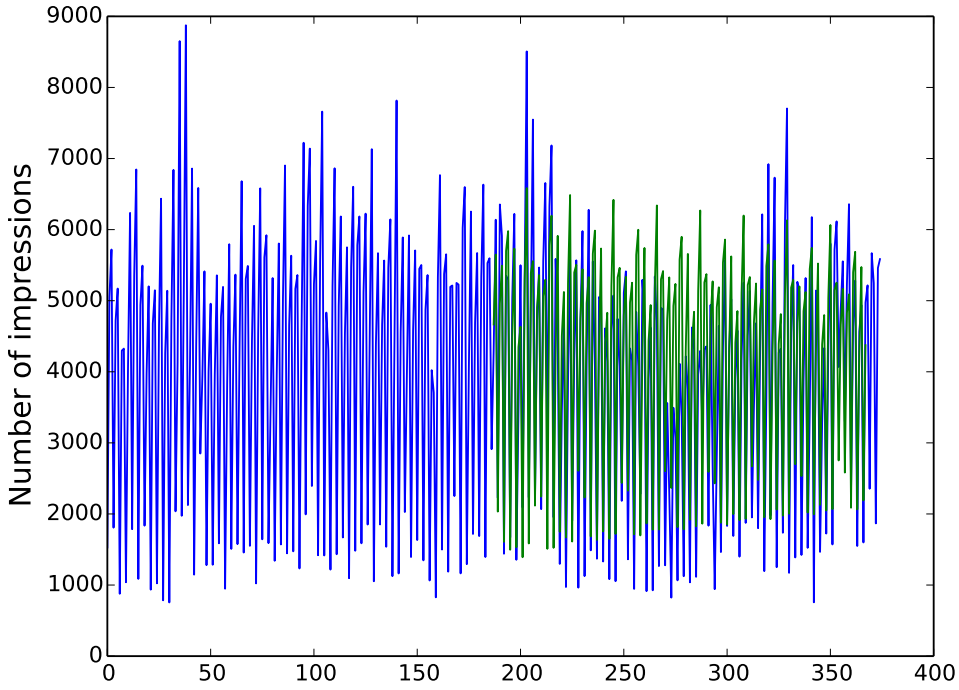

Figure A.36: Arima Allow Drift True - Impressions forecast from 2013-11-26 08:00:00 to 2014-01-25 08:00:00

σ (Real Data)	RMSE	MASE
1785.84	955.35	0.2473

Table A.36: Arima Allow Drift True - Error for Impressions forecast from 2013-11-26 08:00:00 to 2014-01-25 08:00:00

Case 1

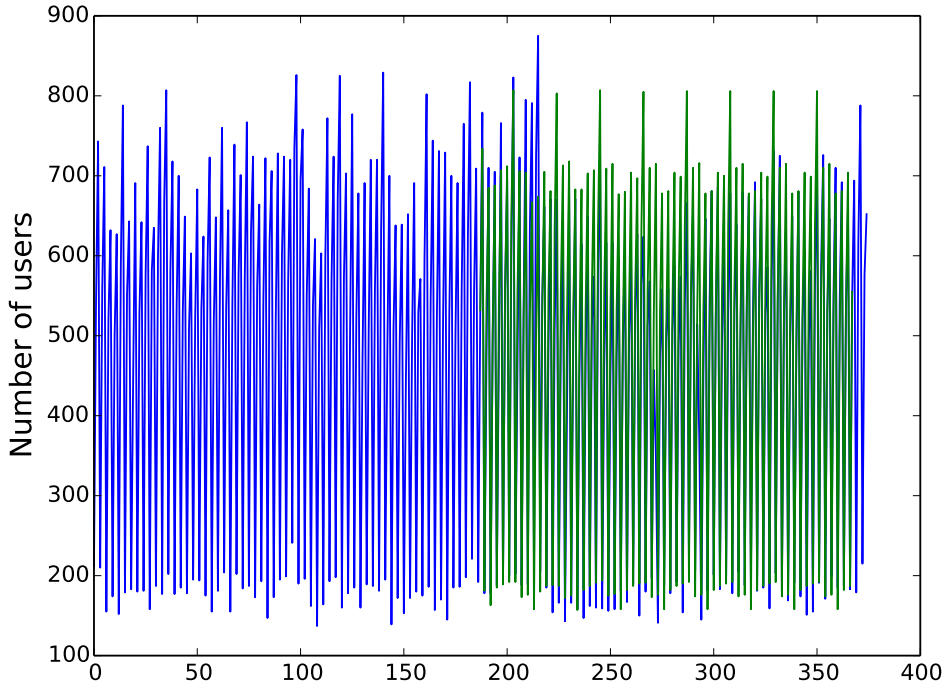

Figure A.37: Arima Allow Drift True - Uniques forecast from 2013-11-26 08:00:00 to 2014-01-25 08:00:00

σ (Real Data)	RMSE	MASE
209.5	70.95	0.1313

Table A.37: Arima Allow Drift True - Error for Uniques forecast from 2013-11-26 08:00:00 to 2014-01-25 08:00:00

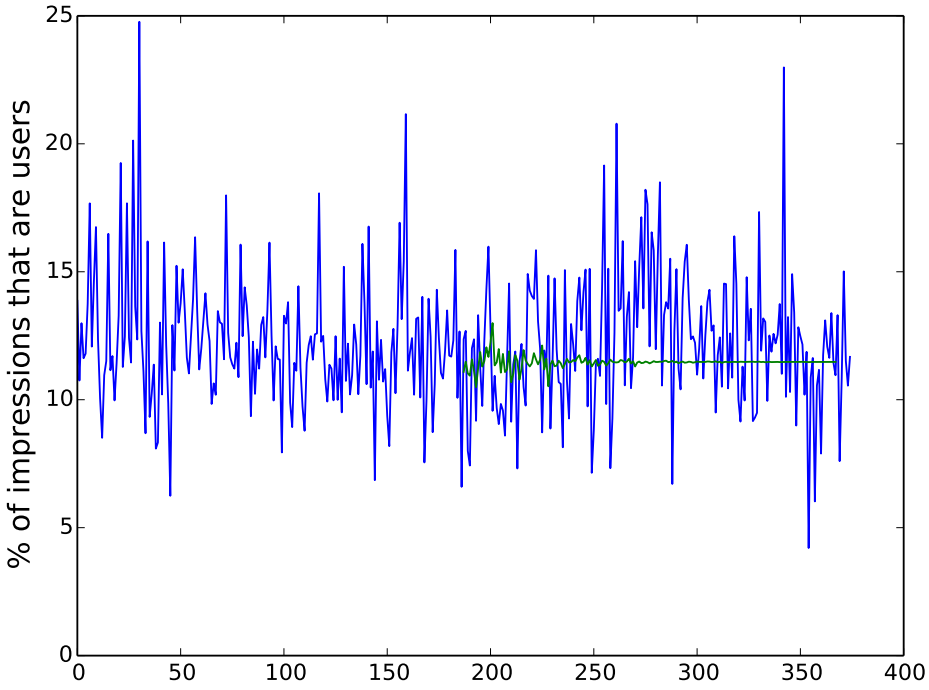

Figure A.38: Arima Allow Drift True - Uniques Percentage forecast from 2013-11-26 08:00:00 to 2014-01-25 08:00:00

σ (Real Data)	RMSE	MASE
2.66	2.79	0.7844

Table A.38: Arima Allow Drift True - Error for Uniques Percentage forecast from 2013-11-26 08:00:00 to 2014-01-25 08:00:00

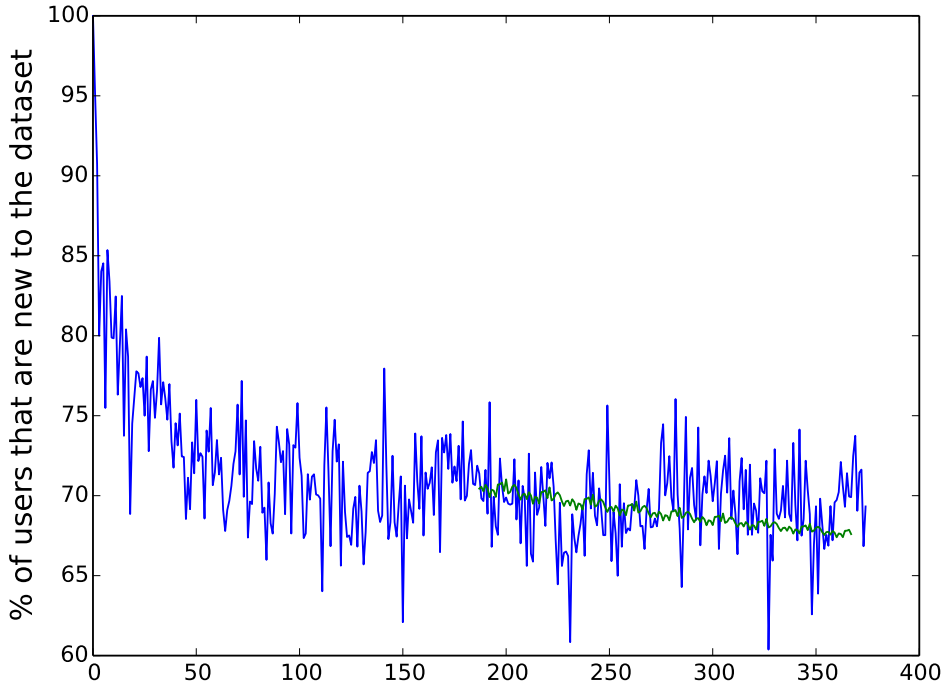

Figure A.39: Arima Allow Drift True - New uniques forecast from 2013-11-26 08:00:00 to 2014-01-25 08:00:00

σ (Real Data)	RMSE	MASE
2.47	2.65	0.6796

Table A.39: Arima Allow Drift True - Error for New Uniques forecast from 2013-11-26 08:00:00 to 2014-01-25 08:00:00

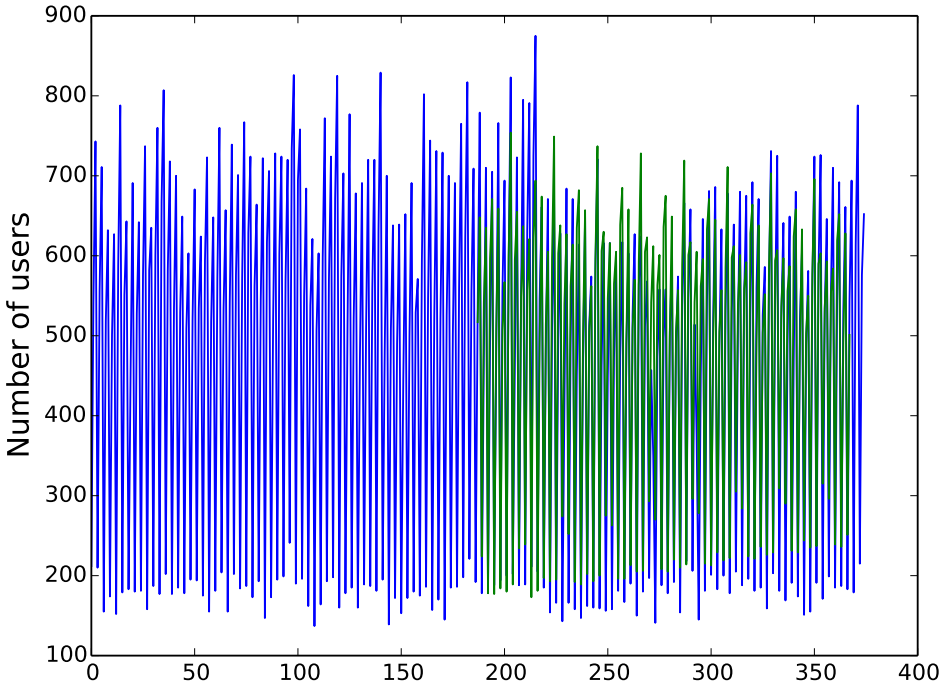

Figure A.40: Arima Allow Drift True - Uniques calculated using percentages forecast from 2013-11-26 08:00:00 to 2014-01-25 08:00:00

σ (Real Data)	RMSE	MASE
209.5	75.08	0.1623

Table A.40: Arima Allow Drift True - Error for Uniques calculated using percentages forecast from 2013-11-26 08:00:00 to 2014-01-25 08:00:00

A.9 Arima Allow Drift False - 8h

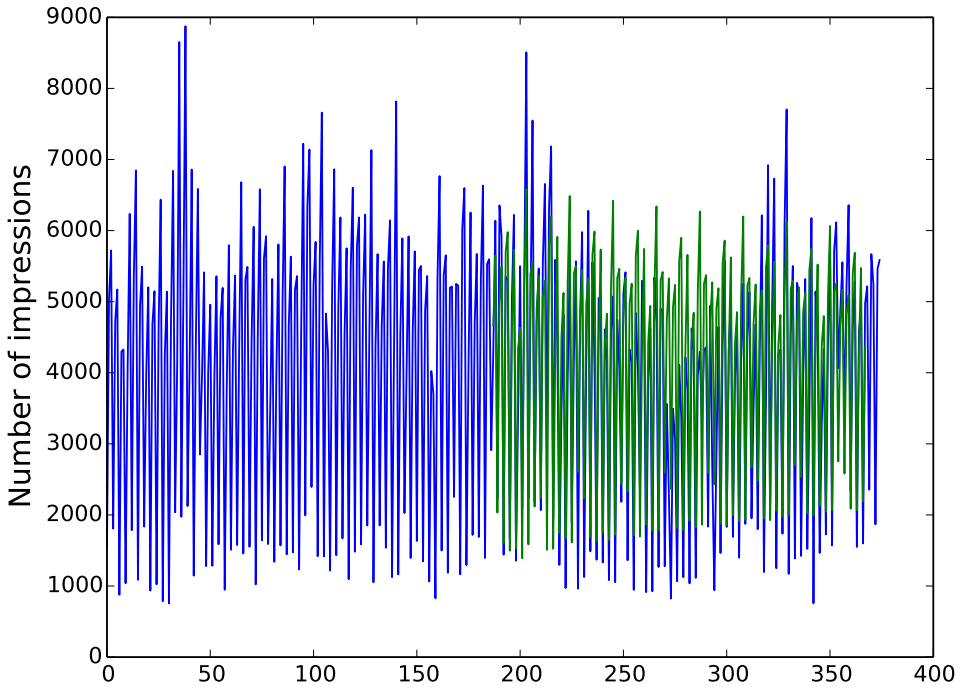

Figure A.41: Arima Allow Drift False - Impressions forecast from 2013-11-26 08:00:00 to 2014-01-25 08:00:00

σ (Real Data)	RMSE	MASE
1785.84	955.35	0.2473

Table A.41: Arima Allow Drift False - Error for Impressions forecast from 2013-11-26 08:00:00 to 2014-01-25 08:00:00

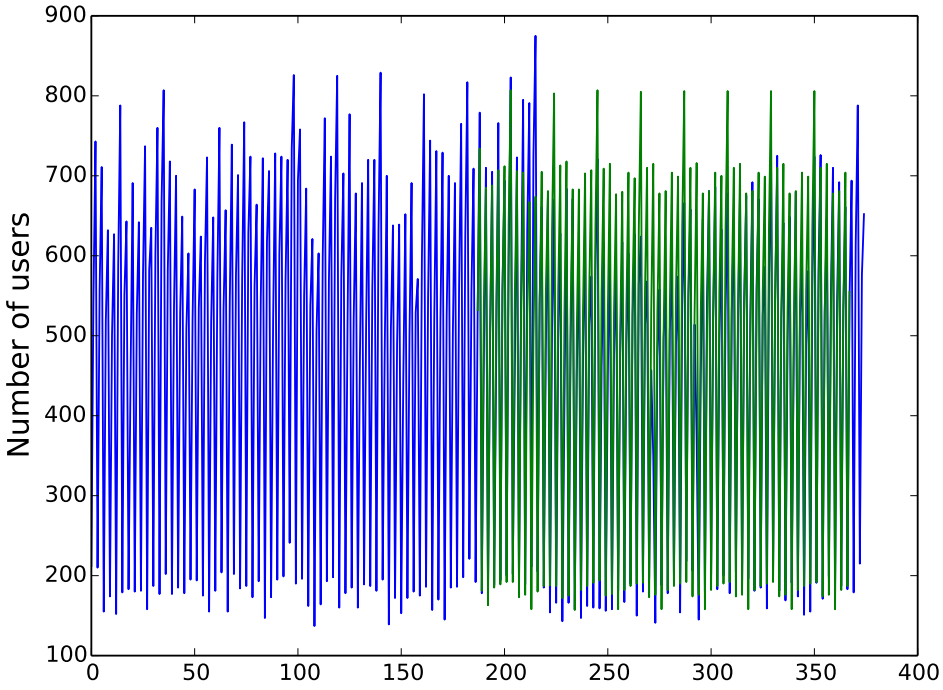

Figure A.42: Arima Allow Drift False - Uniques forecast from 2013-11-26 08:00:00 to 2014-01-25 08:00:00

σ (Real Data)	RMSE	MASE
209.5	70.95	0.1313

Table A.42: Arima Allow Drift False - Error for Uniques forecast from 2013-11-26 08:00:00 to 2014-01-25 08:00:00

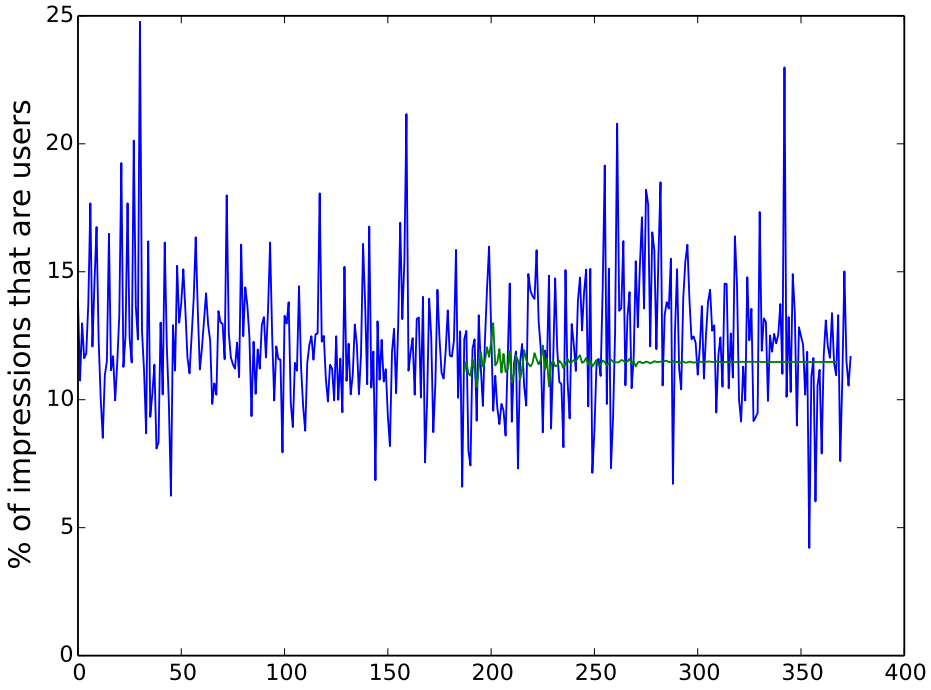

Figure A.43: Arima Allow Drift False - Uniques Percentage forecast from 2013-11-26 08:00:00 to 2014-01-25 08:00:00

σ (Real Data)	RMSE	MASE
2.66	2.79	0.7844

Table A.43: Arima Allow Drift False - Error for Uniques Percentage forecast from 2013-11-26 08:00:00 to 2014-01-25 08:00:00

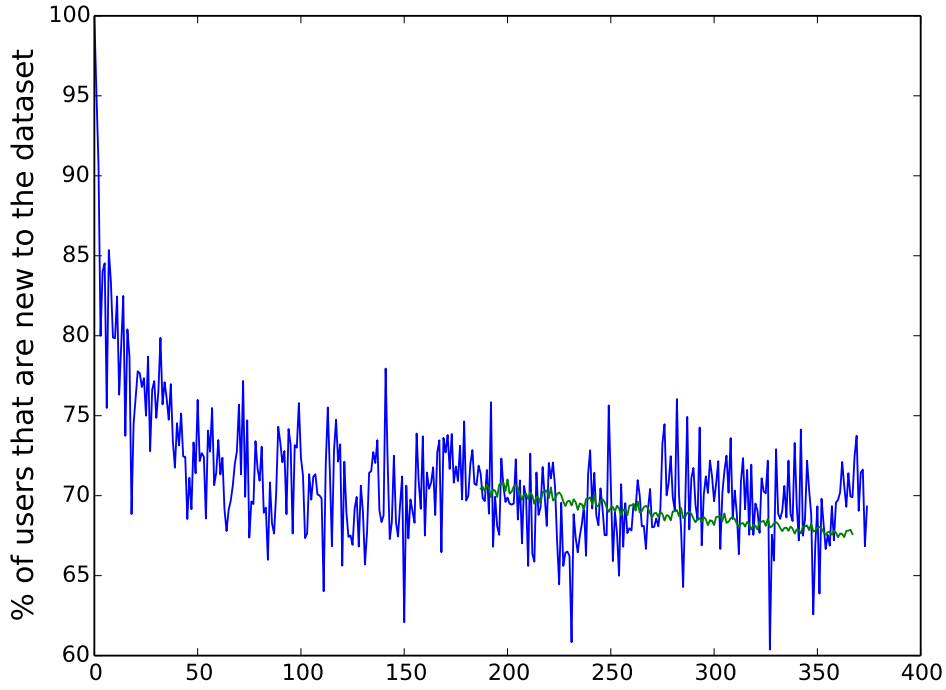

Figure A.44: Arima Allow Drift False - New uniques forecast from 2013-11-26 08:00:00 to 2014-01-25 08:00:00

σ (Real Data)	RMSE	MASE
2.47	2.65	0.6796

Table A.44: Arima Allow Drift False - Error for New Uniques forecast from 2013-11-26 08:00:00 to 2014-01-25 08:00:00

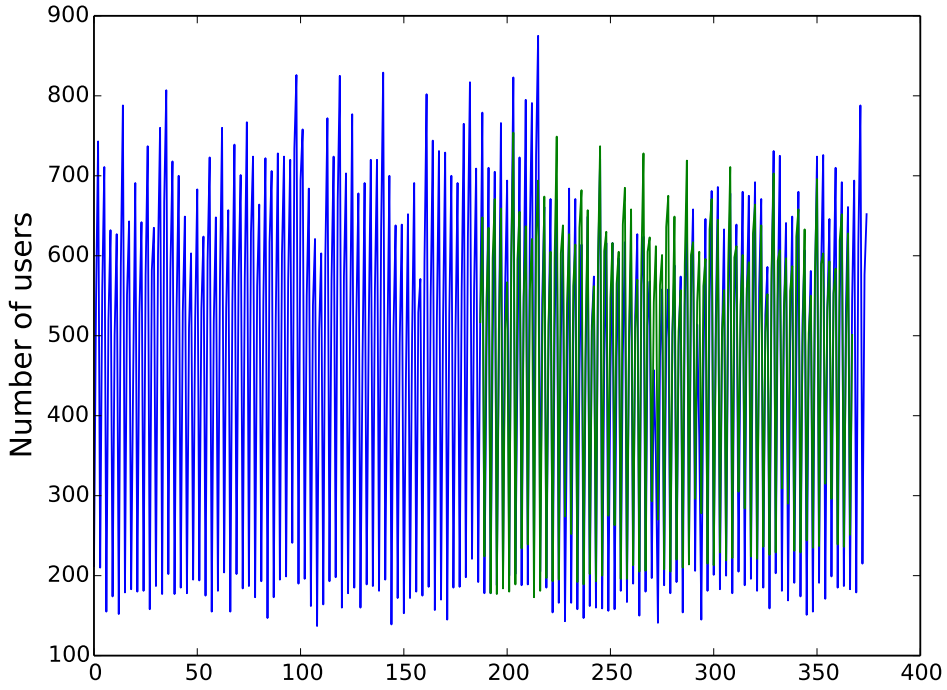

Figure A.45: Arima Allow Drift False - Uniques calculated using percentages forecast from 2013-11-26 08:00:00 to 2014-01-25 08:00:00

σ (Real Data)	RMSE	MASE
209.5	75.08	0.1623

Table A.45: Arima Allow Drift False - Error for Uniques calculated using percentages forecast from 2013-11-26 08:00:00 to 2014-01-25 08:00:00

A.10 Baseline - 12h

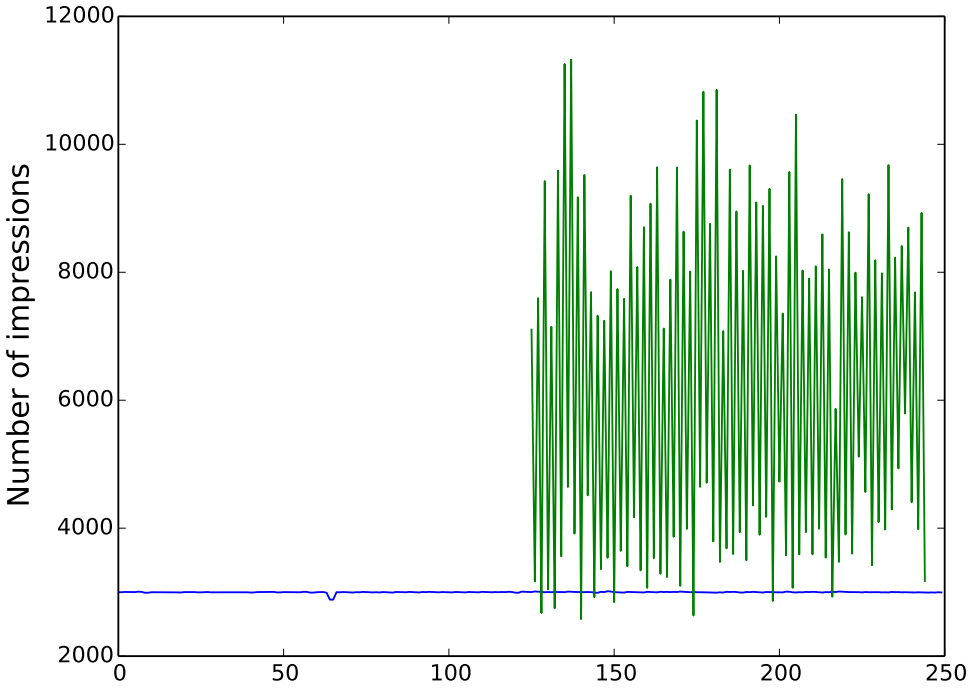

Figure A.46: Baseline - Impressions forecast from 2013-11-26 12:00:00 to 2014-01-25 00:00:00

σ (Real Data)	RMSE	MASE
4.33	4120.76	628.6889

Table A.46: Baseline - Error for Impressions forecast from 2013-11-26 12:00:00 to 2014-01-25 00:00:00

Case 1

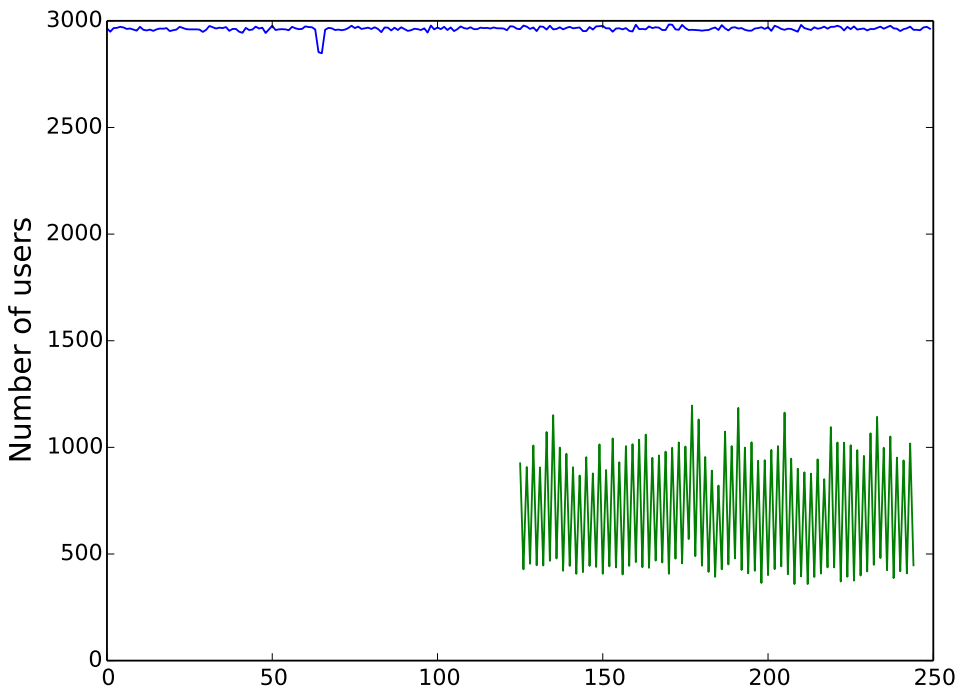

Figure A.47: Baseline - Uniques forecast from 2013-11-26 12:00:00 to 2014-01-25 00:00:00

σ (Real Data)	RMSE	MASE
7.86	2273.11	226.2229

Table A.47: Baseline - Error for Uniques forecast from 2013-11-26 12:00:00 to 2014-01-25 00:00:00

Case 1

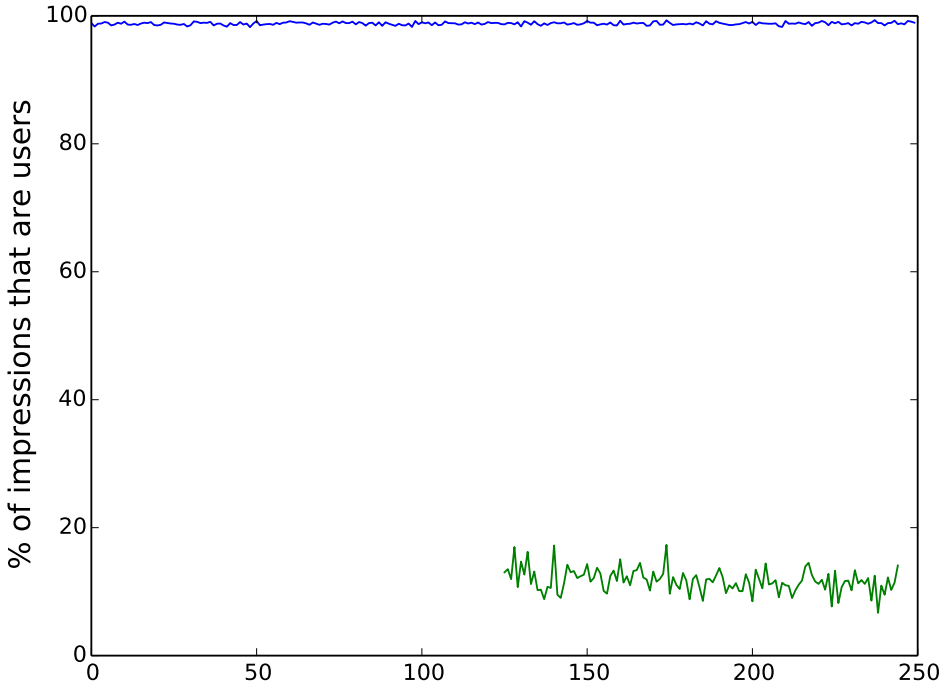

Figure A.48: Baseline - Uniques Percentage forecast from 2013-11-26 12:00:00 to 2014-01-25 00:00:00

σ (Real Data)	RMSE	MASE
0.22	87.08	345.7132

Table A.48: Baseline - Error for Uniques Percentage forecast from 2013-11-26 12:00:00 to 2014-01-25 00:00:00

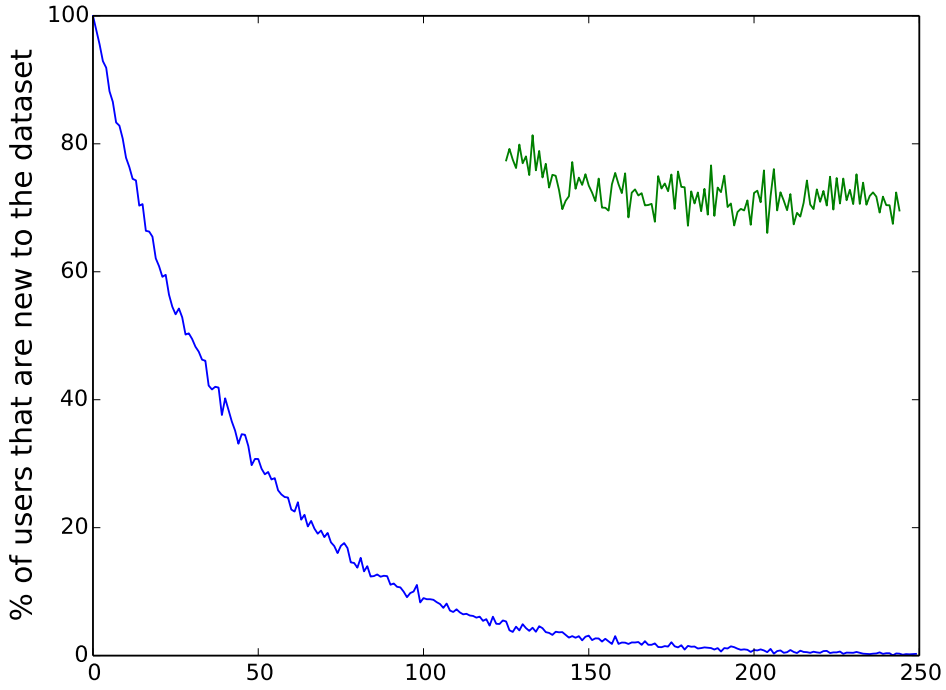

Figure A.49: Baseline - New uniques forecast from 2013-11-26 12:00:00 to 2014-01-25 00:00:00

σ (Real Data)	RMSE	MASE
1.29	70.81	60.1692

Table A.49: Baseline - Error for New Uniques forecast from 2013-11-26 12:00:00 to 2014-01-25 00:00:00

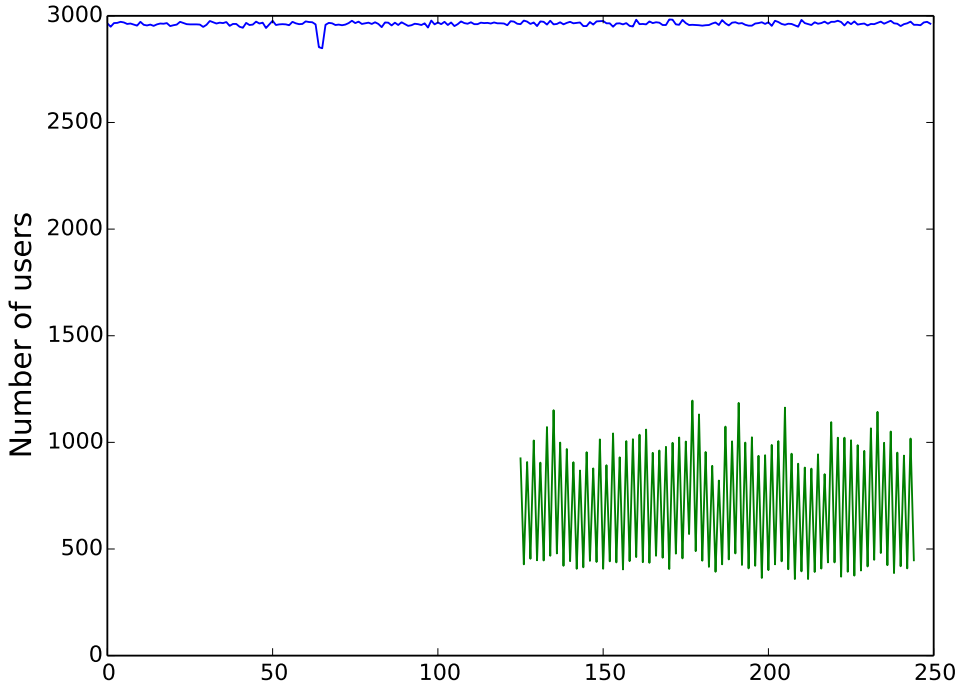

Figure A.50: Baseline - Uniques calculated using percentages forecast from 2013-11-26 12:00:00 to 2014-01-25 00:00:00

σ (Real Data)	RMSE	MASE
7.86	2273.25	226.2371

Table A.50: Baseline - Error for Uniques calculated using percentages forecast from 2013-11-26 12:00:00 to 2014-01-25 00:00:00

A.11 Arima Allow Drift True - 12h

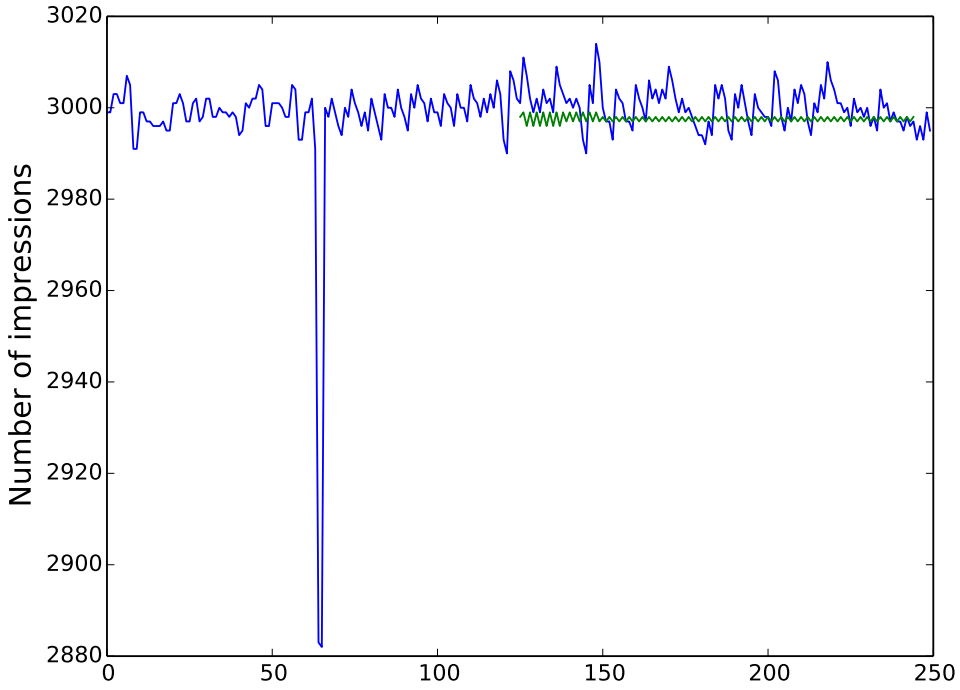

Figure A.51: Arima Allow Drift True - Impressions forecast from 2013-11-26 12:00:00 to 2014-01-25 00:00:00

σ (Real Data)	RMSE	MASE
4.33	4.98	0.7711

Table A.51: Arima Allow Drift True - Error for Impressions forecast from 2013-11-26 12:00:00 to 2014-01-25 00:00:00

Case 1

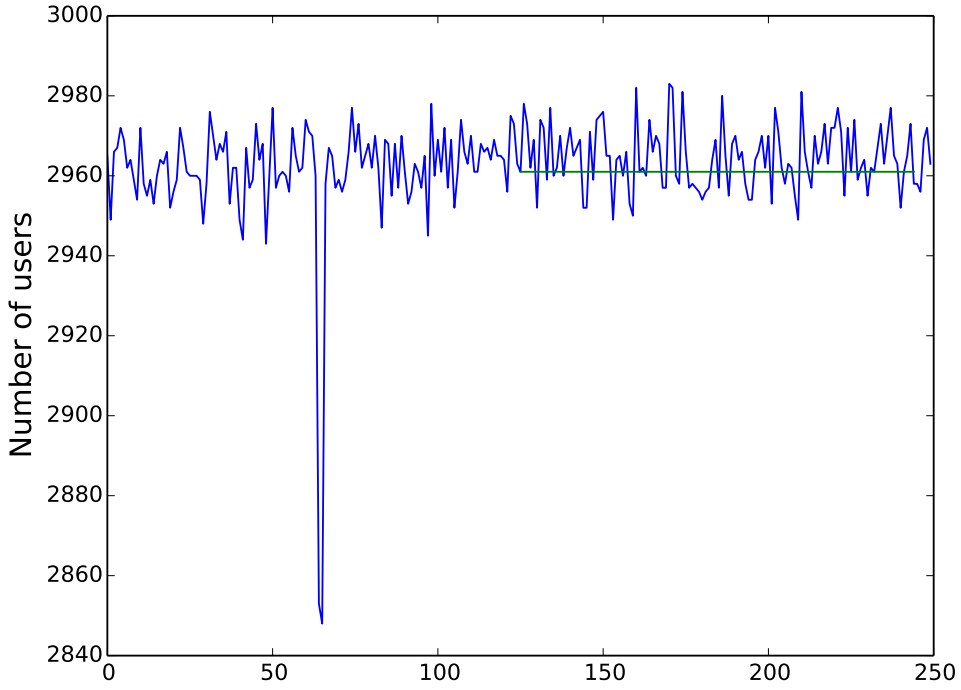

Figure A.52: Arima Allow Drift True - Uniques forecast from 2013-11-26 12:00:00 to 2014-01-25 00:00:00

σ (Real Data)	RMSE	MASE
7.86	8.72	0.6932

Table A.52: Arima Allow Drift True - Error for Uniques forecast from 2013-11-26 12:00:00 to 2014-01-25 00:00:00

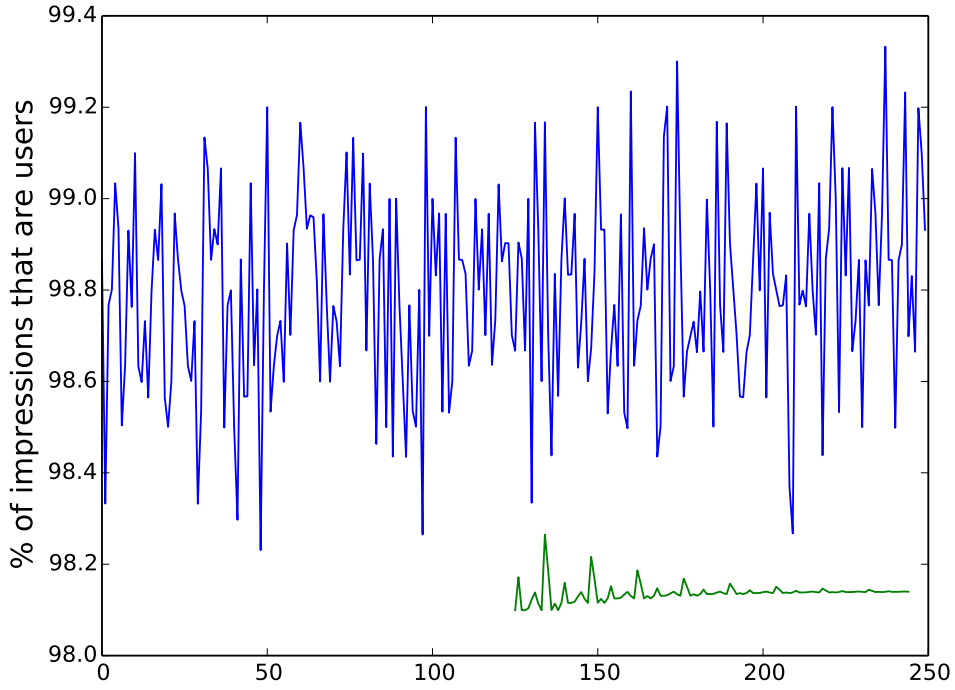

Figure A.53: Arima Allow Drift True - Uniques Percentage forecast from 2013-11-26 12:00:00 to 2014-01-25 00:00:00

σ (Real Data)	RMSE	MASE
0.22	0.71	2.6661

Table A.53: Arima Allow Drift True - Error for Uniques Percentage forecast from 2013-11-26 12:00:00 to 2014-01-25 00:00:00

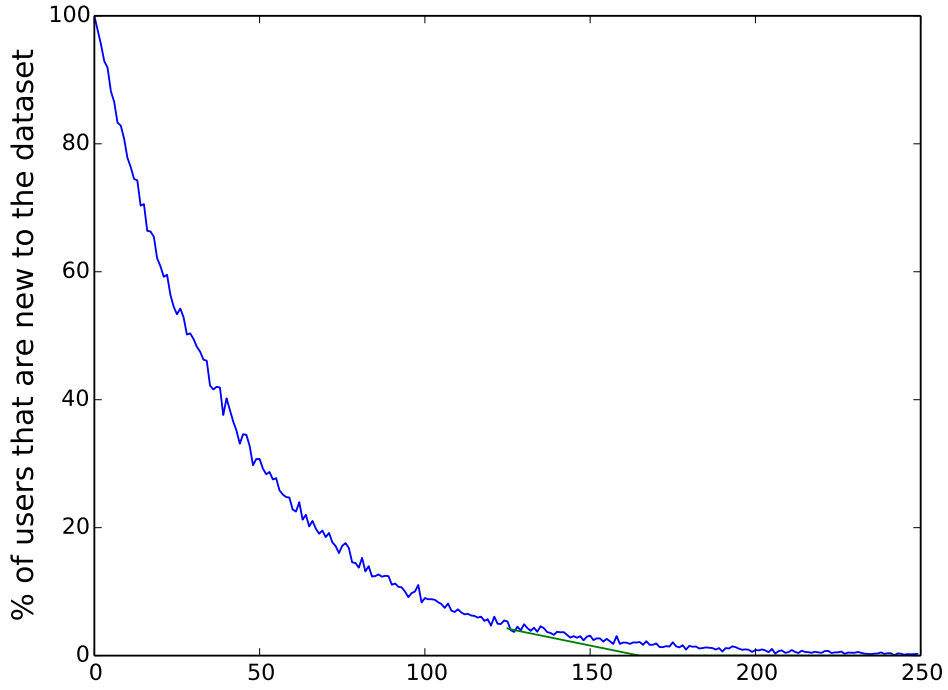

Figure A.54: Arima Allow Drift True - New uniques forecast from 2013-11-26 12:00:00 to 2014-01-25 00:00:00

σ (Real Data)	RMSE	MASE
1.29	1.07	0.8013

Table A.54: Arima Allow Drift True - Error for New Uniques forecast from 2013-11-26 12:00:00 to 2014-01-25 00:00:00

Case 1

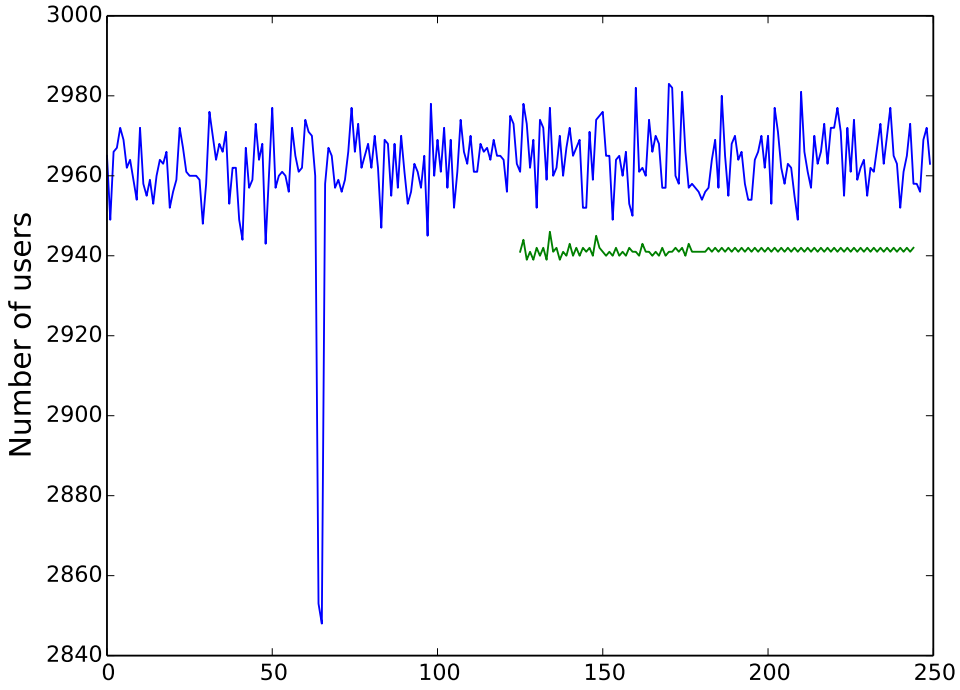

Figure A.55: Arima Allow Drift True - Uniques calculated using percentages forecast from 2013-11-26 12:00:00 to 2014-01-25 00:00:00

σ (Real Data)	RMSE	MASE
7.86	24.59	2.3387

Table A.55: Arima Allow Drift True - Error for Uniques calculated using percentages forecast from 2013-11-26 12:00:00 to 2014-01-25 00:00:00

A.12 Arima Allow Drift False - 12h

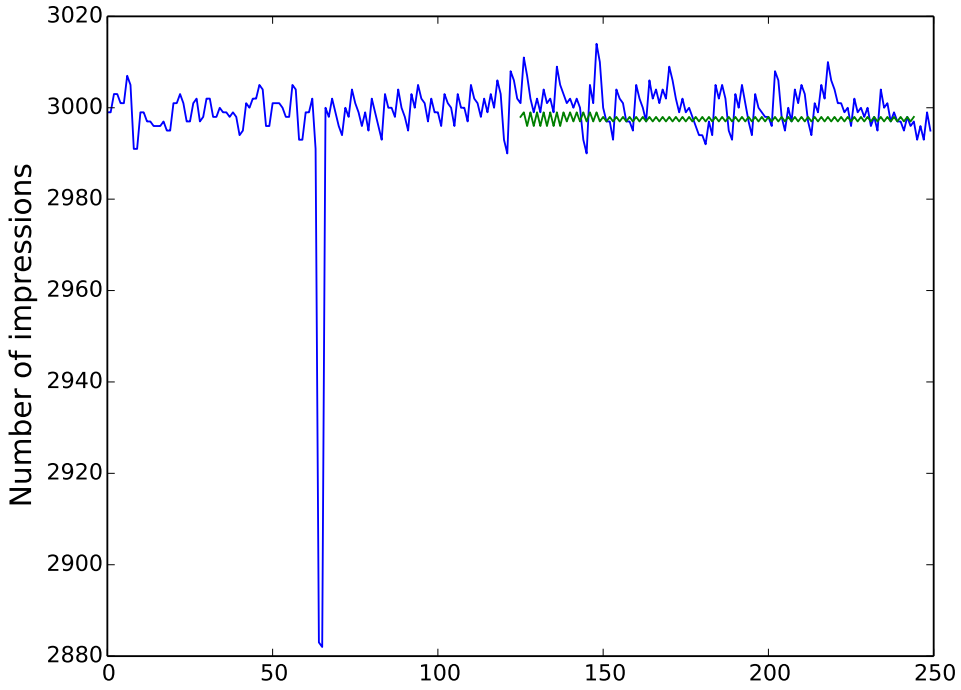

Figure A.56: Arima Allow Drift False - Impressions forecast from 2013-11-26 12:00:00 to 2014-01-25 00:00:00

σ (Real Data)	RMSE	MASE
4.33	4.98	0.7711

Table A.56: Arima Allow Drift False - Error for Impressions forecast from 2013-11-26 12:00:00 to 2014-01-25 00:00:00

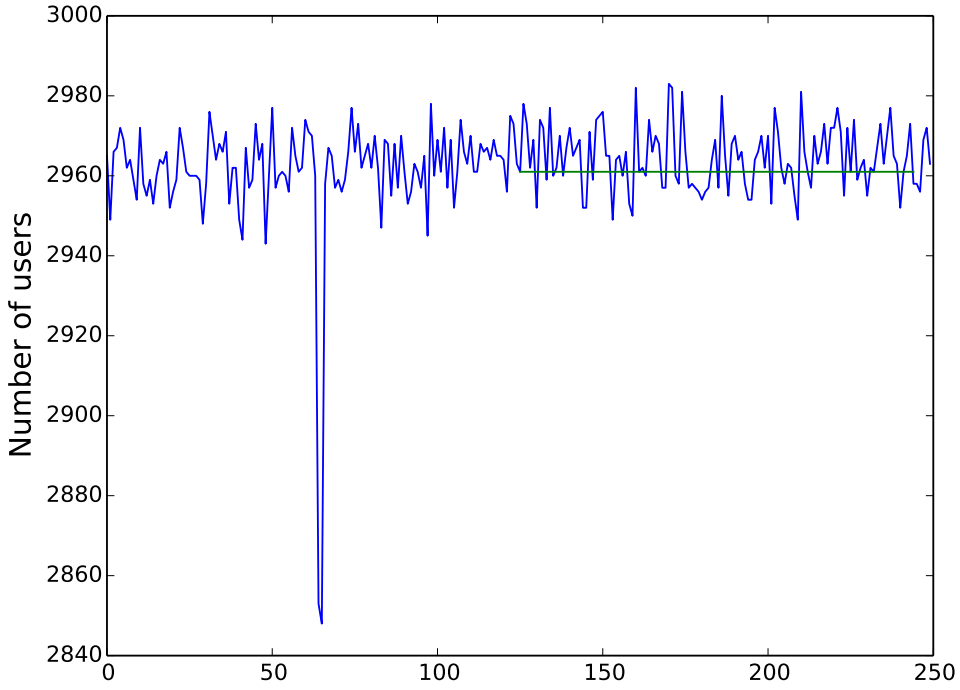

Figure A.57: Arima Allow Drift False - Uniques forecast from 2013-11-26 12:00:00 to 2014-01-25 00:00:00

σ (Real Data)	RMSE	MASE
7.86	8.72	0.6932

Table A.57: Arima Allow Drift False - Error for Uniques forecast from 2013-11-26 12:00:00 to 2014-01-25 00:00:00

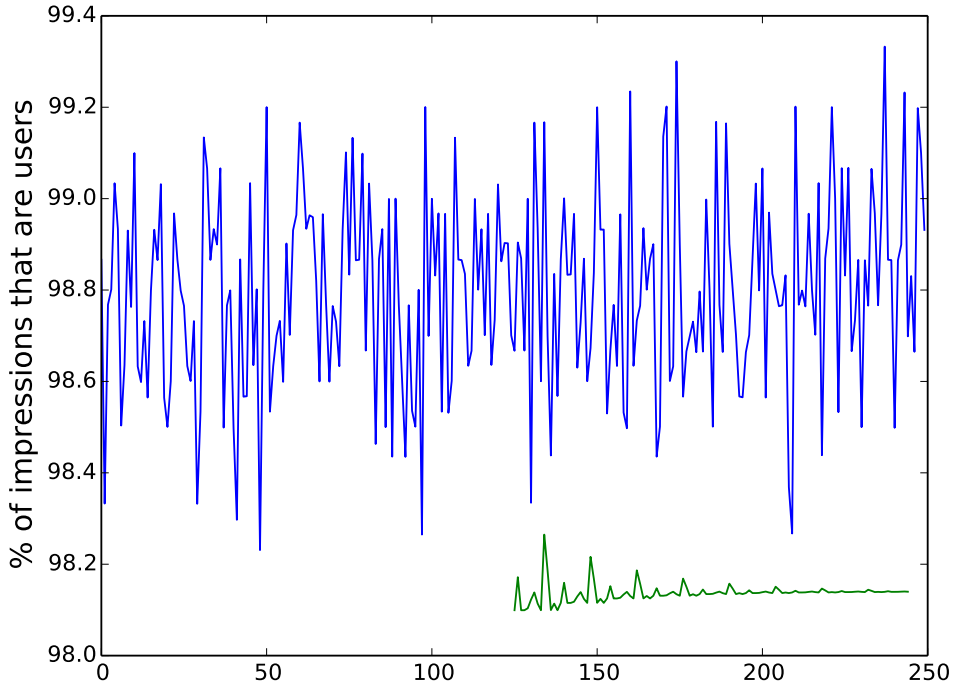

Figure A.58: Arima Allow Drift False - Uniques Percentage forecast from 2013-11-26 12:00:00 to 2014-01-25 00:00:00

σ (Real Data)	RMSE	MASE
0.22	0.71	2.6661

Table A.58: Arima Allow Drift False - Error for Uniques Percentage forecast from 2013-11-26 12:00:00 to 2014-01-25 00:00:00

Case 1

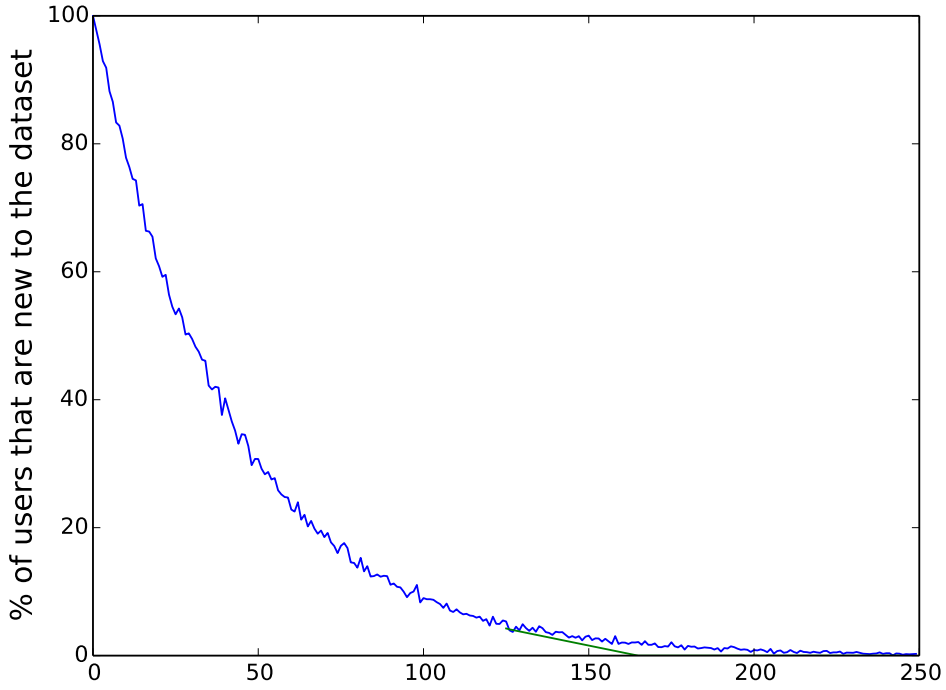

Figure A.59: Arima Allow Drift False - New uniques forecast from 2013-11-26 12:00:00 to 2014-01-25 00:00:00

σ (Real Data)	RMSE	MASE
1.29	1.07	0.8013

Table A.59: Arima Allow Drift False - Error for New Uniques forecast from 2013-11-26 12:00:00 to 2014-01-25 00:00:00

Case 1

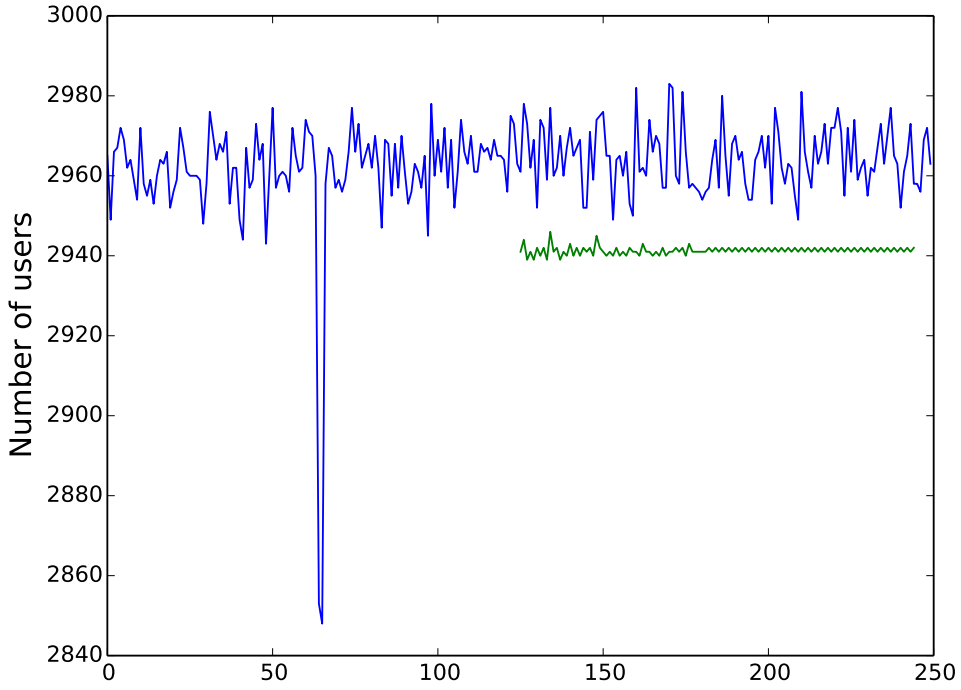

Figure A.60: Arima Allow Drift False - Uniques calculated using percentages forecast from 2013-11-26 12:00:00 to 2014-01-25 00:00:00

σ (Real Data)	RMSE	MASE
7.86	24.59	2.3387

Table A.60: Arima Allow Drift False - Error for Uniques calculated using percentages forecast from 2013-11-26 12:00:00 to 2014-01-25 00:00:00

A.13 Baseline - 24h

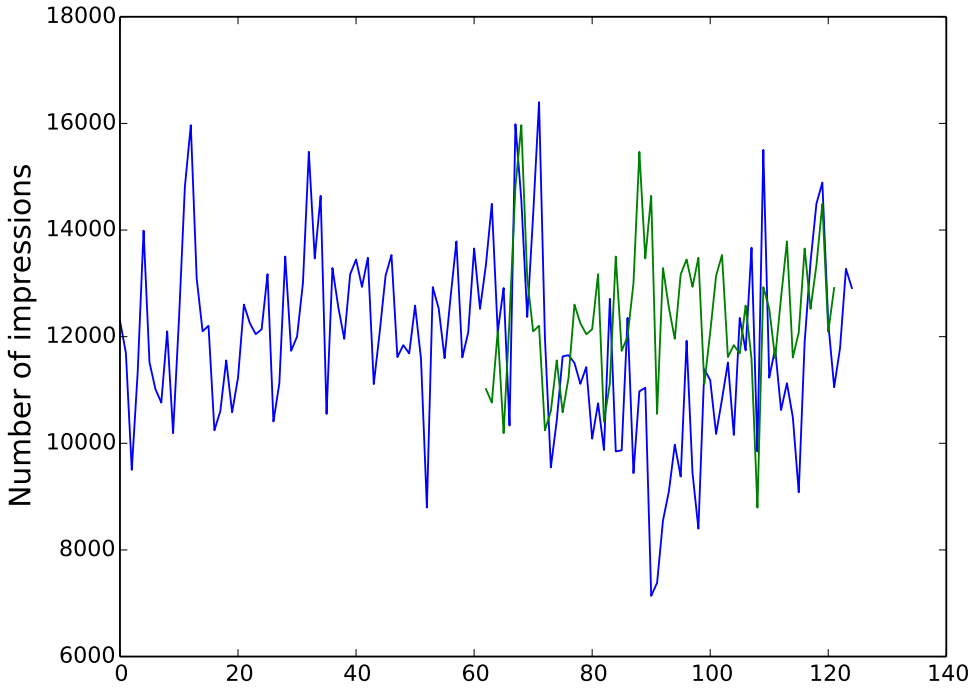

Figure A.61: Baseline - Impressions forecast from 2013-11-26 00:00:00 to 2014-01-24 00:00:00

σ (Real Data)	RMSE	MASE
1946.7	2448.97	1.4338

Table A.61: Baseline - Error for Impressions forecast from 2013-11-26 00:00:00 to 2014-01-24 00:00:00

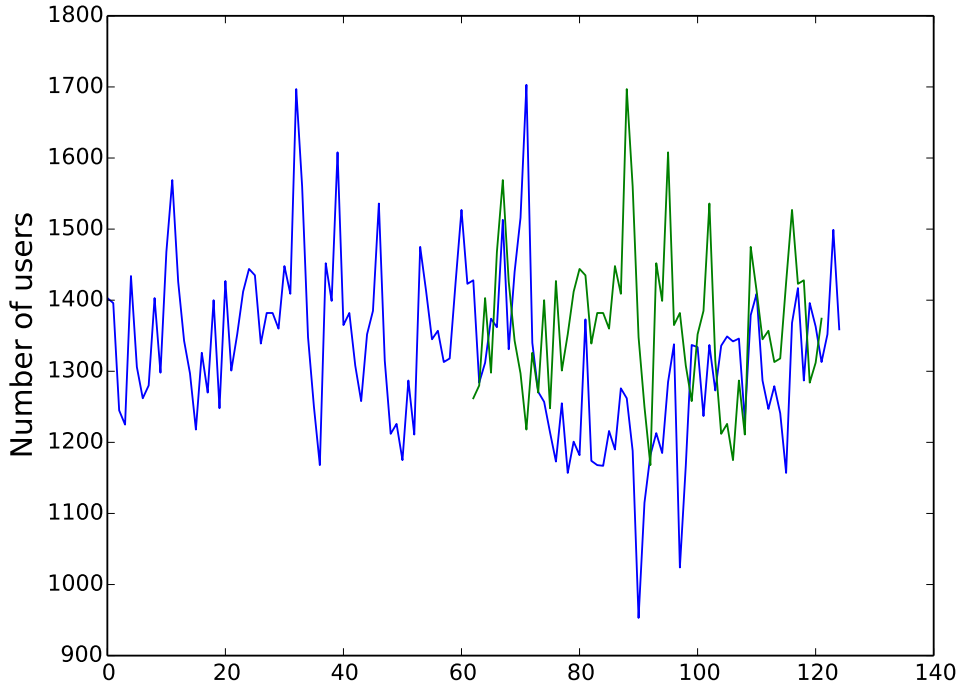

Figure A.62: Baseline - Uniques forecast from 2013-11-26 00:00:00 to 2014-01-24 00:00:00

σ (Real Data)	RMSE	MASE
118.7	181.8	1.4049

Table A.62: Baseline - Error for Uniques forecast from 2013-11-26 00:00:00 to 2014-01-24 00:00:00

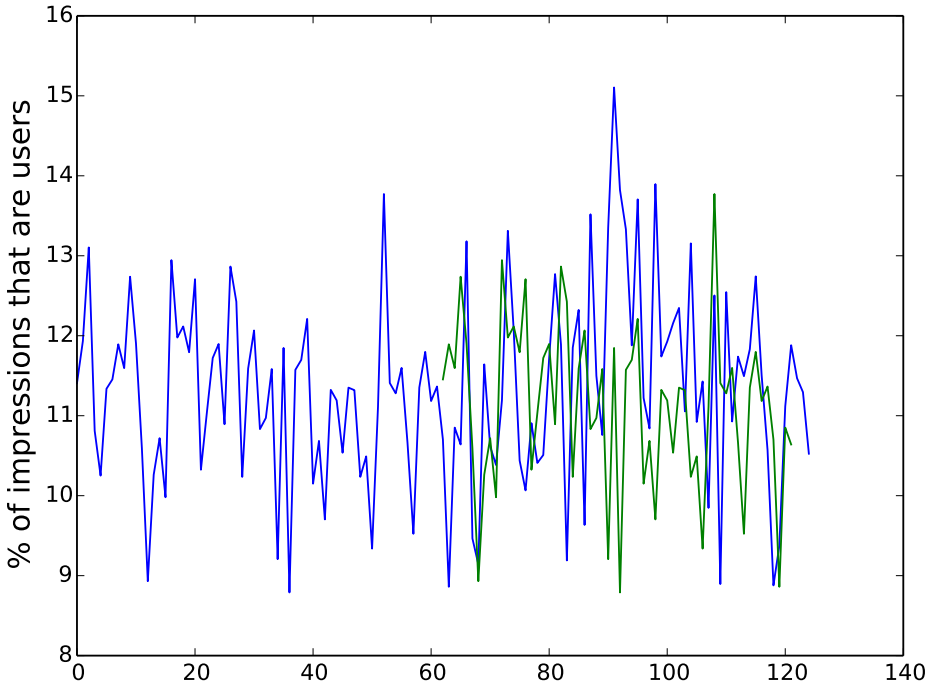

Figure A.63: Baseline - Uniques Percentage forecast from 2013-11-26 00:00:00 to 2014-01-24 00:00:00

σ (Real Data)	RMSE	MASE
1.36	1.76	1.2386

Table A.63: Baseline - Error for Uniques Percentage forecast from 2013-11-26 00:00:00 to 2014-01-24 00:00:00

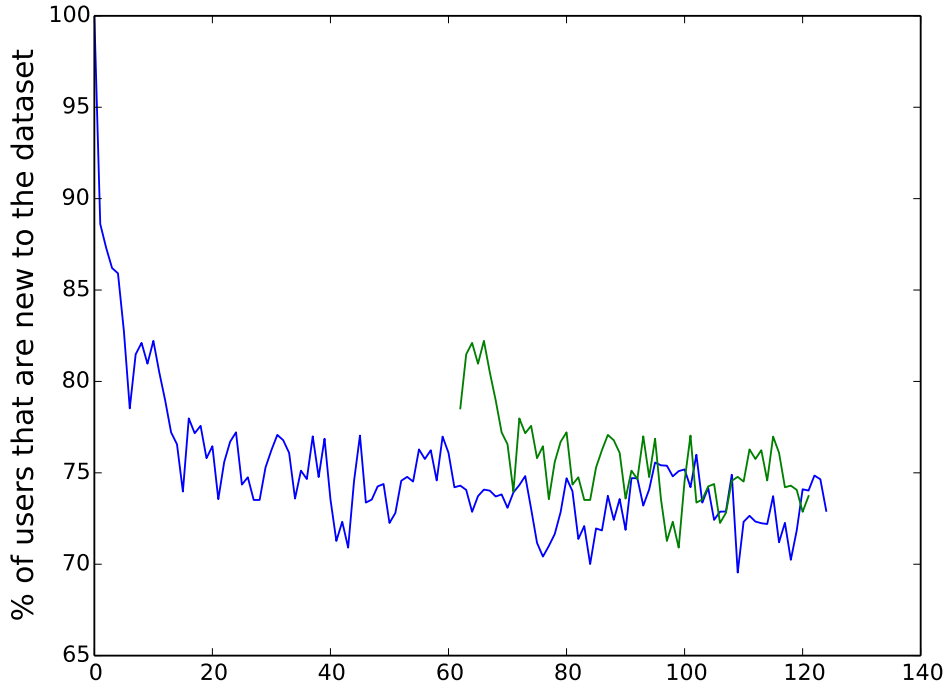

Figure A.64: Baseline - New uniques forecast from 2013-11-26 00:00:00 to 2014-01-24 00:00:00

σ (Real Data)	RMSE	MASE
1.47	3.71	1.7509

Table A.64: Baseline - Error for New Uniques forecast from 2013-11-26 00:00:00 to 2014-01-24 00:00:00

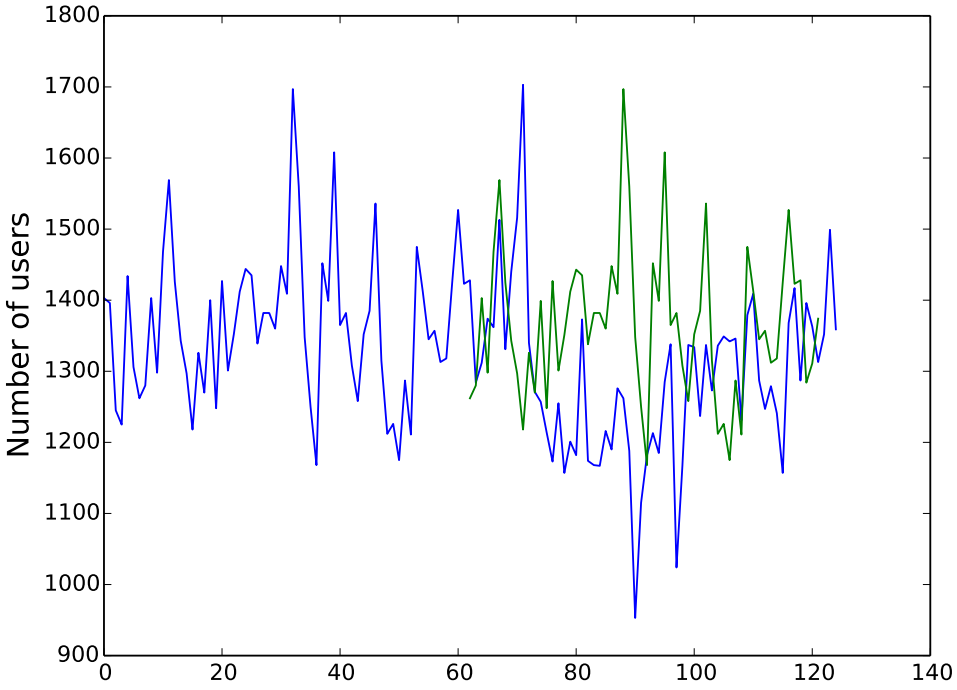

Figure A.65: Baseline - Uniques calculated using percentages forecast from 2013-11-26 00:00:00 to 2014-01-24 00:00:00

σ (Real Data)	RMSE	MASE
118.7	181.73	1.4041

Table A.65: Baseline - Error for Uniques calculated using percentages forecast from 2013-11-26 00:00:00 to 2014-01-24 00:00:00

A.14 Arima Allow Drift True - 24h

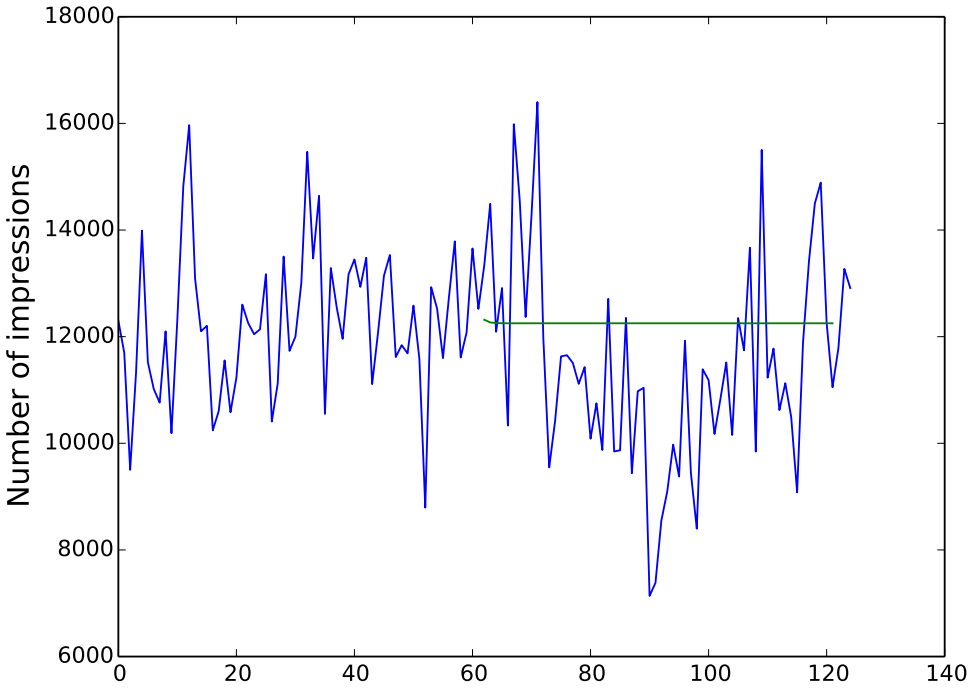

Figure A.66: Arima Allow Drift True - Impressions forecast from 2013-11-26 00:00:00 to 2014-01-24 00:00:00

σ (Real Data)	RMSE	MASE
1946.7	2139.99	1.2725

Table A.66: Arima Allow Drift True - Error for Impressions forecast from 2013-11-26 00:00:00 to 2014-01-24 00:00:00

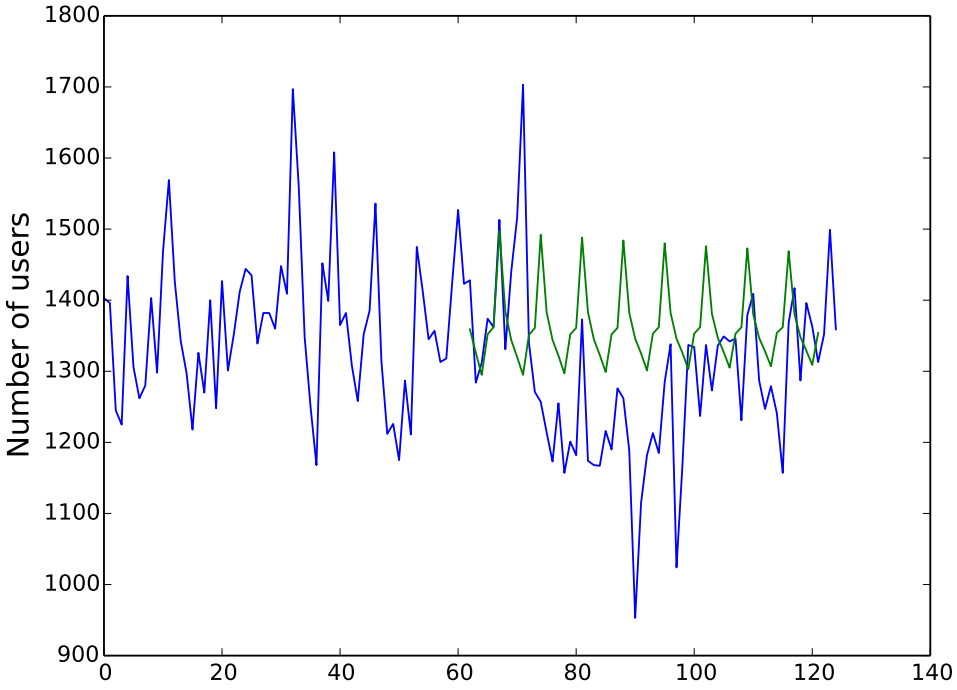

Figure A.67: Arima Allow Drift True - Uniques forecast from 2013-11-26 00:00:00 to 2014-01-24 00:00:00

σ (Real Data)	RMSE	MASE
118.7	145.15	1.1328

Table A.67: Arima Allow Drift True - Error for Uniques forecast from 2013-11-26 00:00:00 to 2014-01-24 00:00:00

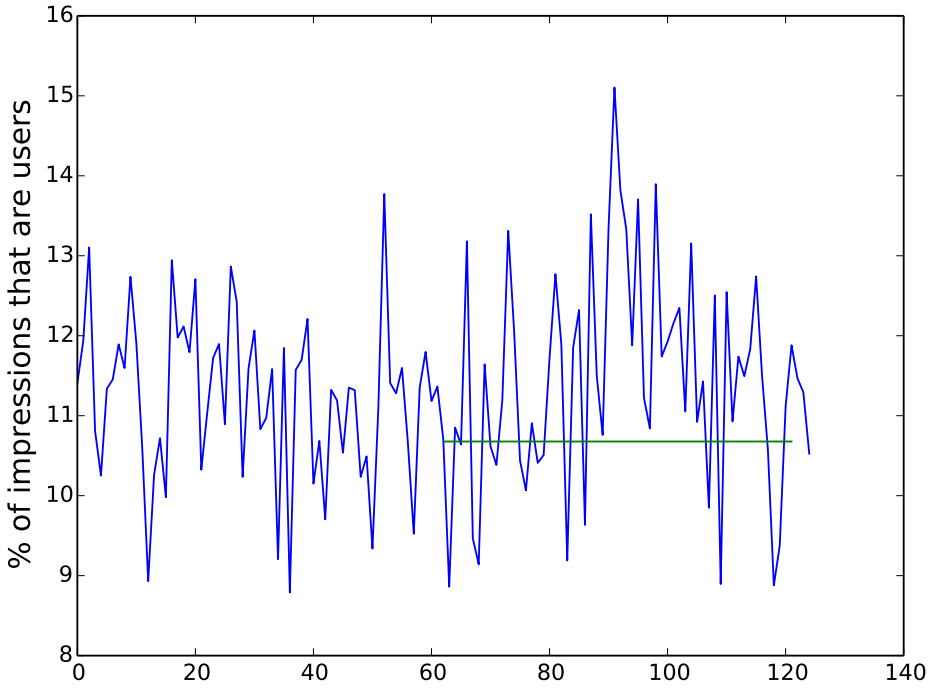

Figure A.68: Arima Allow Drift True - Uniques Percentage forecast from 2013-11-26 00:00:00 to 2014-01-24 00:00:00

σ (Real Data)	RMSE	MASE
1.36	1.59	1.132

Table A.68: Arima Allow Drift True - Error for Uniques Percentage forecast from 2013-11-26 00:00:00 to 2014-01-24 00:00:00

Case 1

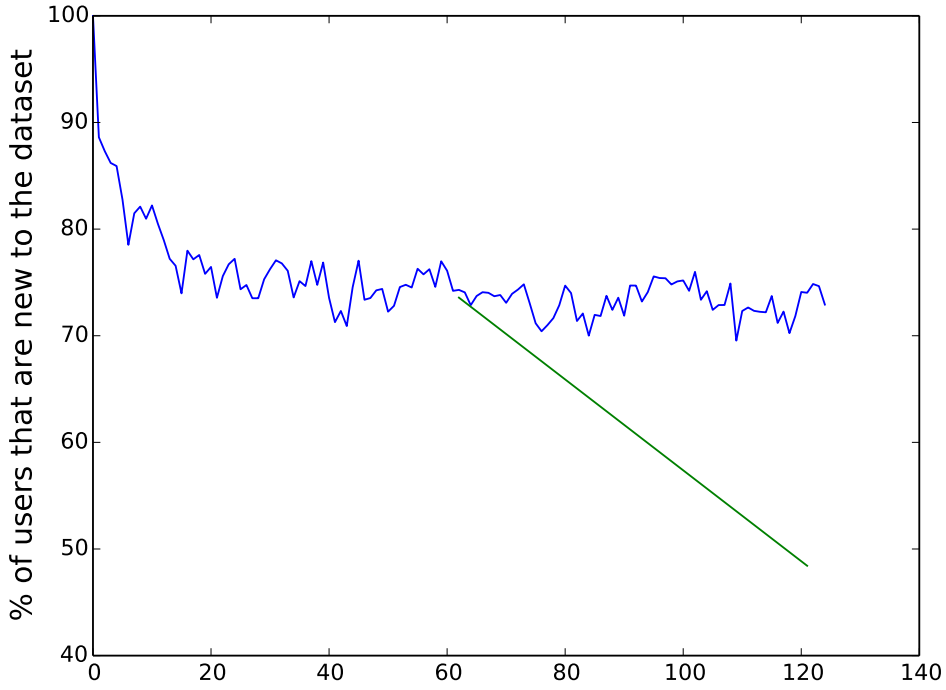

Figure A.69: Arima Allow Drift True - New uniques forecast from 2013-11-26 00:00:00 to 2014-01-24 00:00:00

σ (Real Data)	RMSE	MASE
1.47	14.26	6.9569

Table A.69: Arima Allow Drift True - Error for New Uniques forecast from 2013-11-26 00:00:00 to 2014-01-24 00:00:00

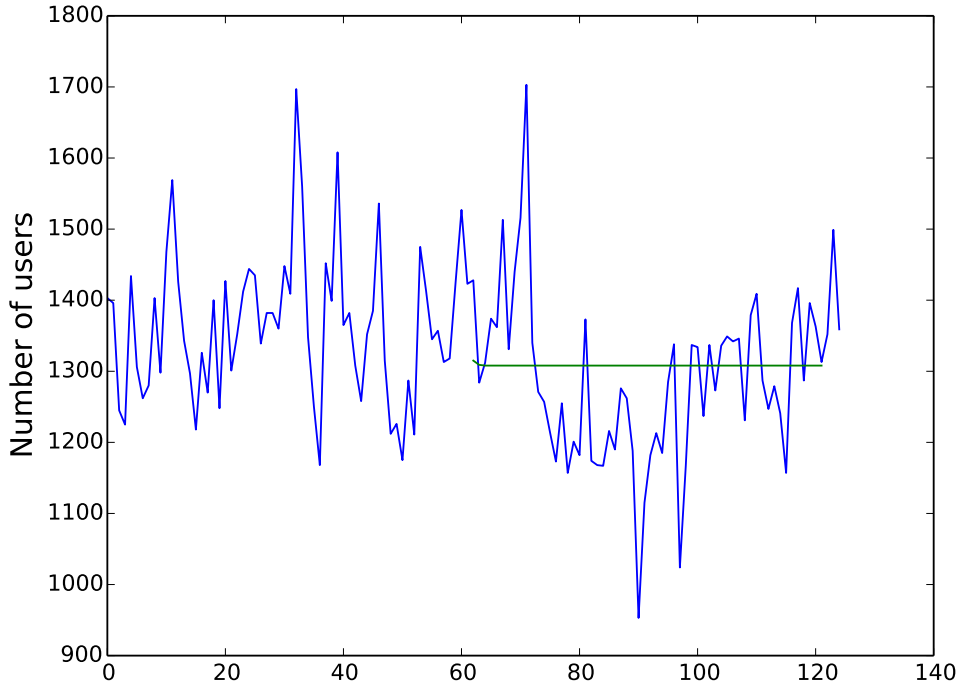

Figure A.70: Arima Allow Drift True - Uniques calculated using percentages forecast from 2013-11-26 00:00:00 to 2014-01-24 00:00:00

σ (Real Data)	RMSE	MASE
118.7	120.12	0.9058

Table A.70: Arima Allow Drift True - Error for Uniques calculated using percentages forecast from 2013-11-26 00:00:00 to 2014-01-24 00:00:00

A.15 Arima Allow Drift False - 24h

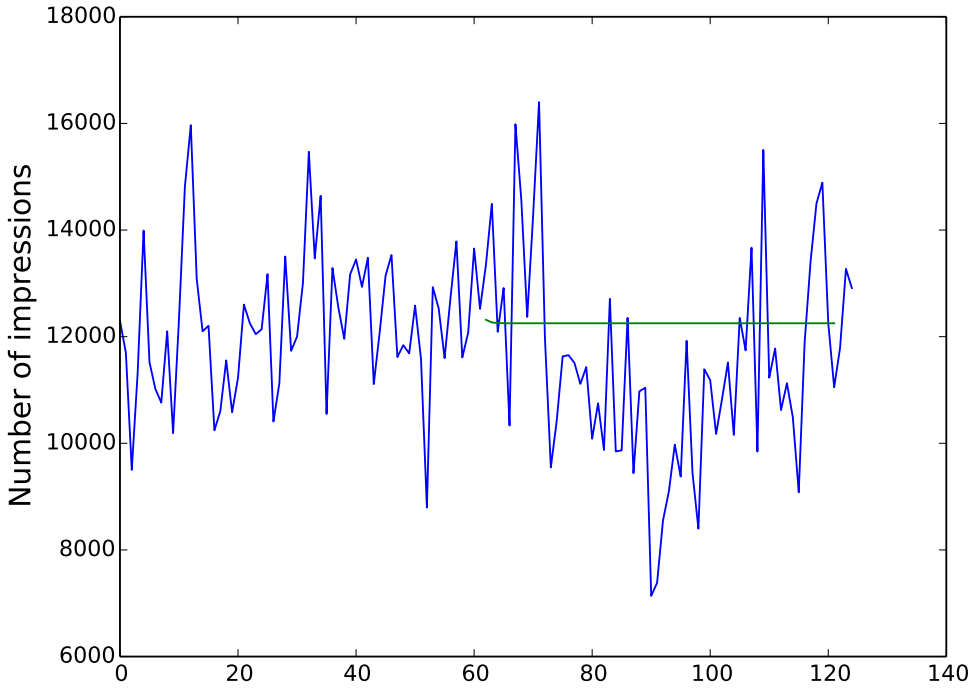

Figure A.71: Arima Allow Drift False - Impressions forecast from 2013-11-26 00:00:00 to 2014-01-24 00:00:00

σ (Real Data)	RMSE	MASE
1946.7	2139.99	1.2725

Table A.71: Arima Allow Drift False - Error for Impressions forecast from 2013-11-26 00:00:00 to 2014-01-24 00:00:00

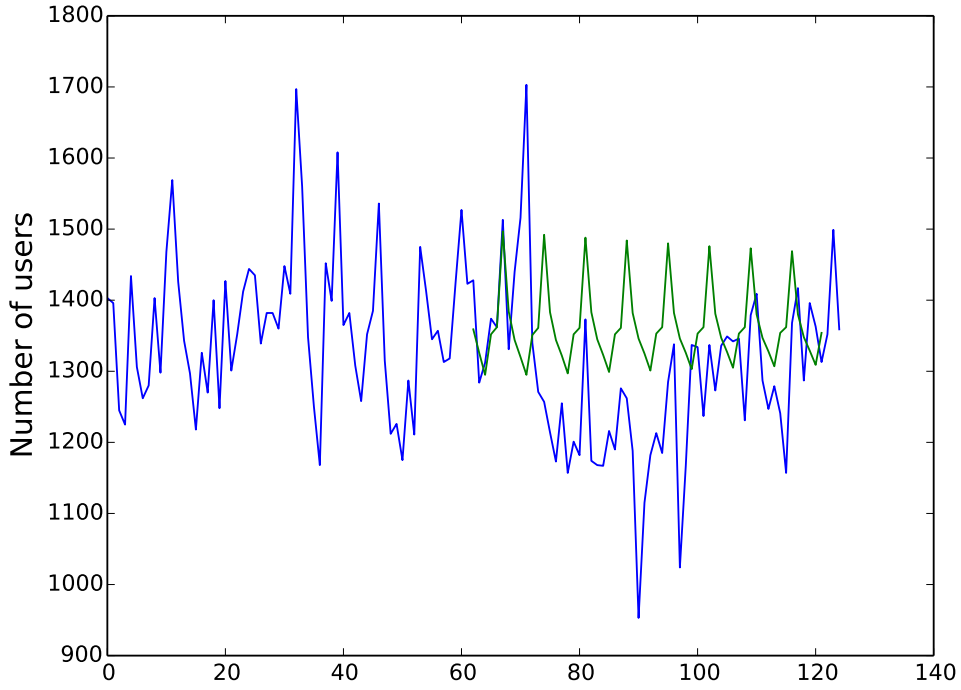

Figure A.72: Arima Allow Drift False - Uniques forecast from 2013-11-26 00:00:00 to 2014-01-24 00:00:00

σ (Real Data)	RMSE	MASE
118.7	145.15	1.1328

Table A.72: Arima Allow Drift False - Error for Uniques forecast from 2013-11-26 00:00:00 to 2014-01-24 00:00:00

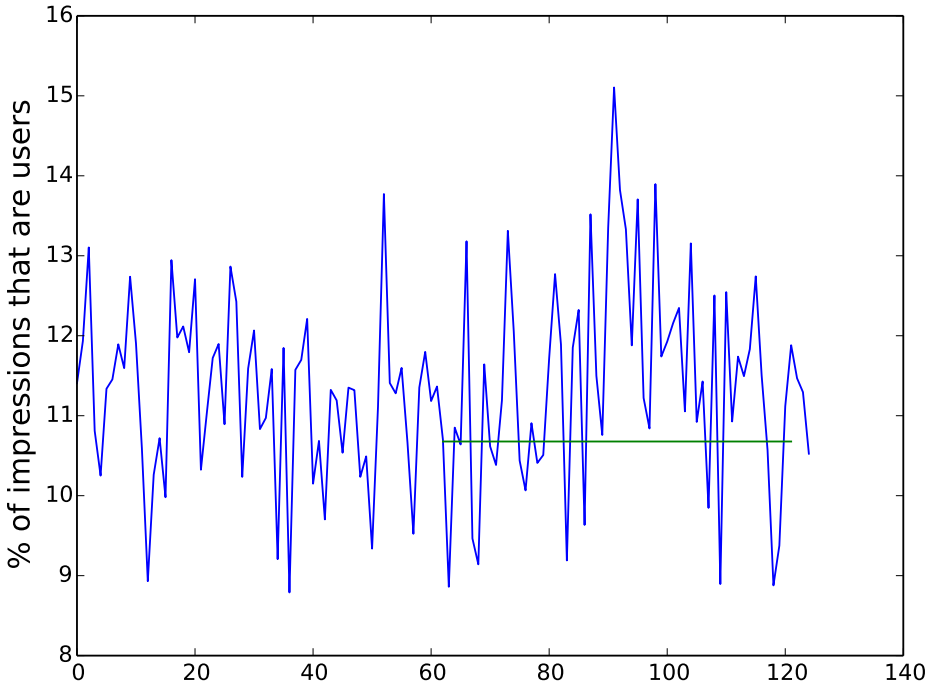

Figure A.73: Arima Allow Drift False - Uniques Percentage forecast from 2013-11-26 00:00:00 to 2014-01-24 00:00:00

σ (Real Data)	RMSE	MASE
1.36	1.59	1.132

Table A.73: Arima Allow Drift False - Error for Uniques Percentage forecast from 2013-11-26 00:00:00 to 2014-01-24 00:00:00

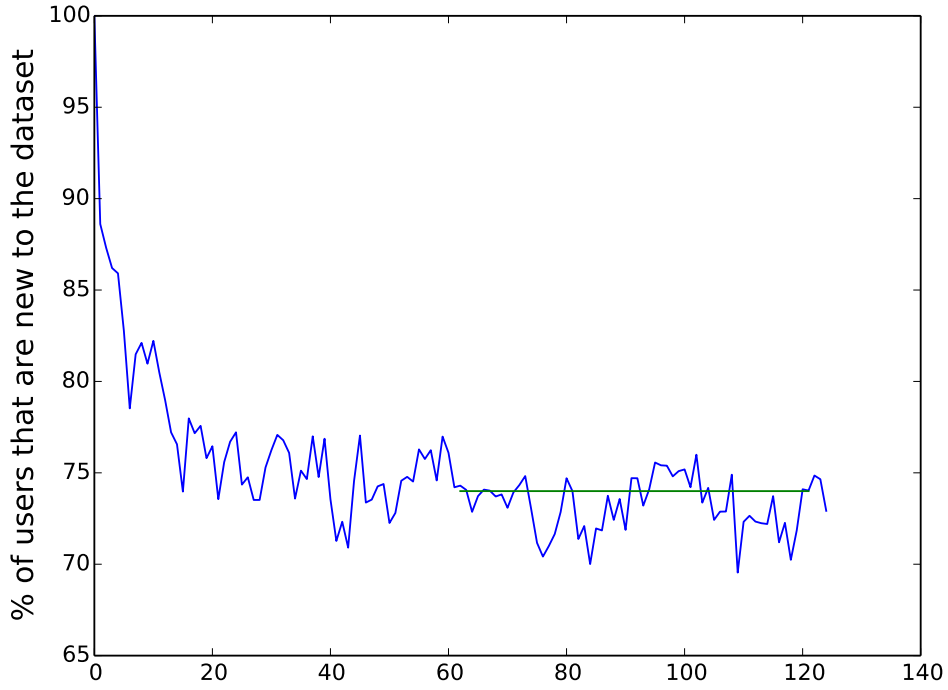

Figure A.74: Arima Allow Drift False - New uniques forecast from 2013-11-26 00:00:00 to 2014-01-24 00:00:00

σ (Real Data)	RMSE	MASE
1.47	1.68	0.7337

Table A.74: Arima Allow Drift False - Error for New Uniques forecast from 2013-11-26 00:00:00 to 2014-01-24 00:00:00

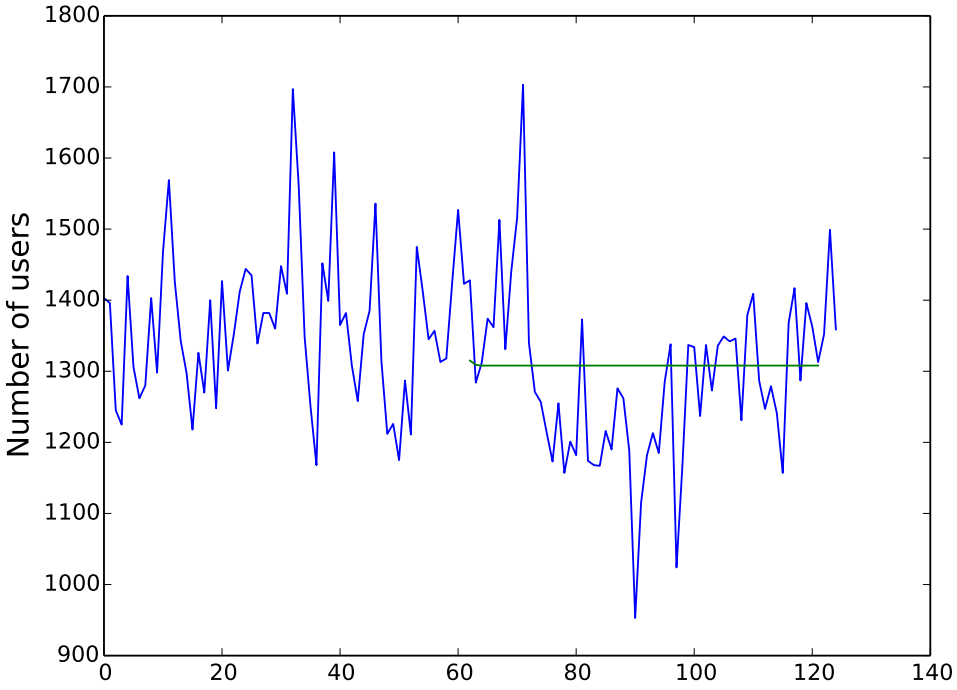

Figure A.75: Arima Allow Drift False - Uniques calculated using percentages forecast from 2013-11-26 00:00:00 to 2014-01-24 00:00:00

σ (Real Data)	RMSE	MASE
118.7	120.12	0.9058

Table A.75: Arima Allow Drift False - Error for Uniques calculated using percentages forecast from 2013-11-26 00:00:00 to 2014-01-24 00:00:00

Appendix B

Case 2

B.1 Baseline - 4h

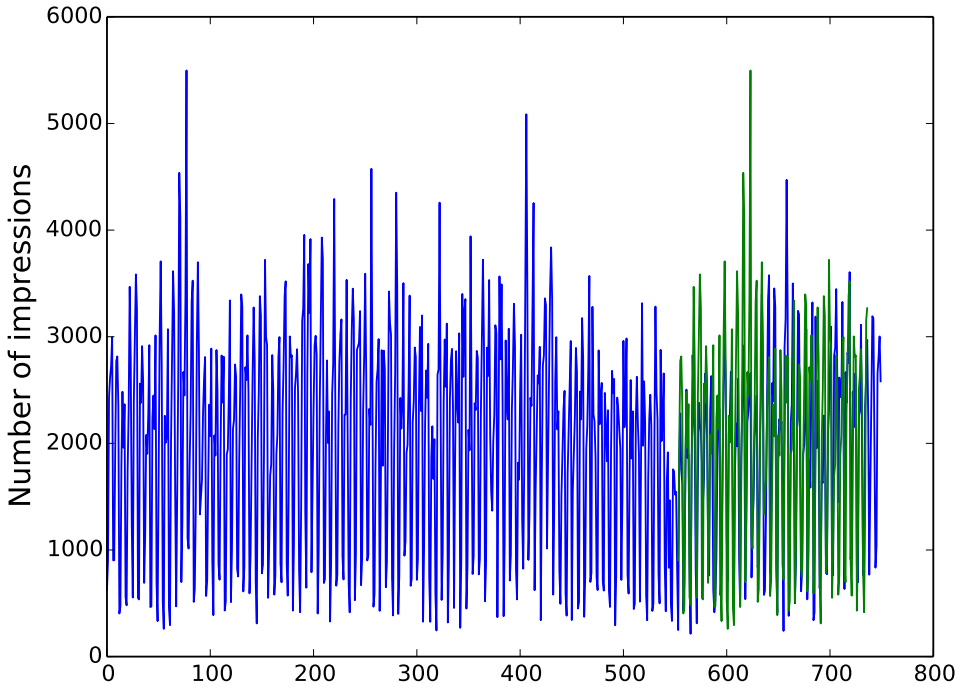

Figure B.1: Baseline - Impressions forecast from 2013-12-26 04:00:00 to 2014-01-25 16:00:00

Case 2

σ (Real Data)	RMSE	MASE
908.7	727.31	0.1922

Table B.1: Baseline - Error for Impressions forecast from 2013-12-26 04:00:00 to 2014-01-25 16:00:00

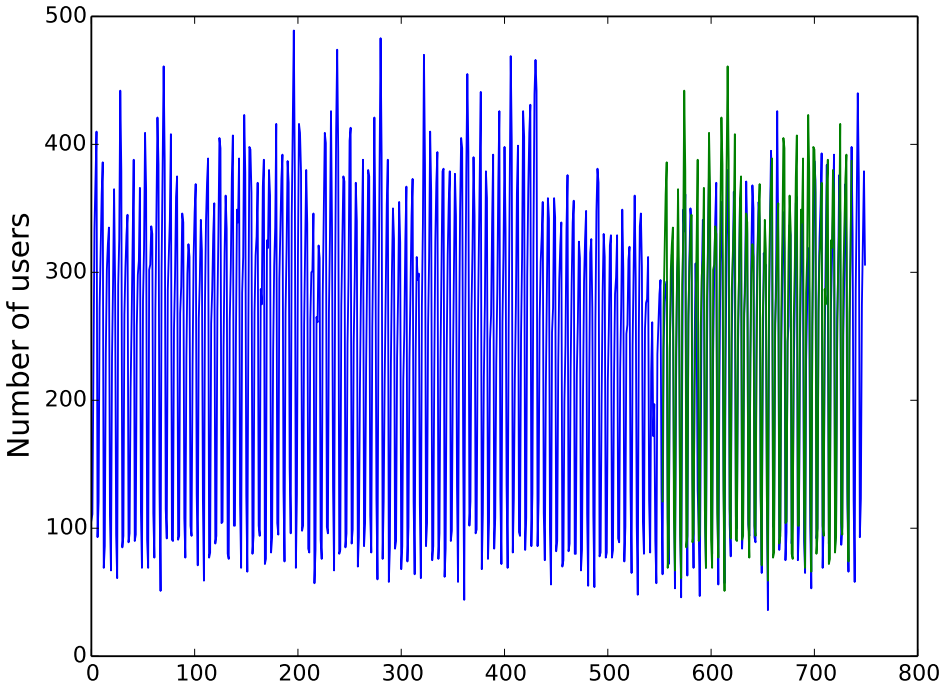

Figure B.2: Baseline - Uniques forecast from 2013-12-26 04:00:00 to 2014-01-25 16:00:00

σ (Real Data)	RMSE	MASE
110.89	43.14	0.0992

Table B.2: Baseline - Error for Uniques forecast from 2013-12-26 04:00:00 to 2014-01-25 16:00:00

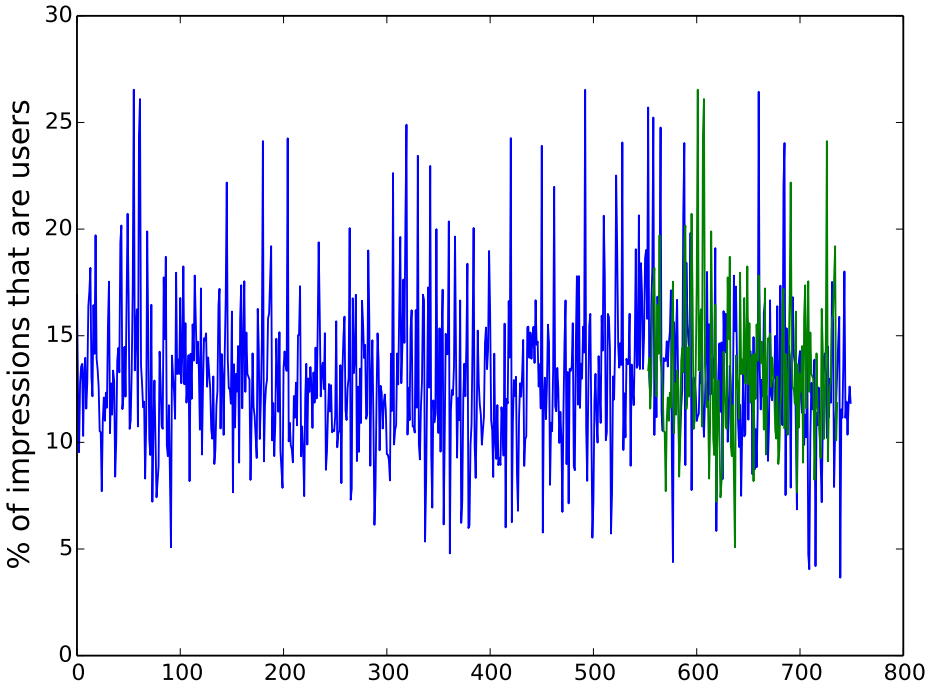

Figure B.3: Baseline - Uniques Percentage forecast from 2013-12-26 04:00:00 to 2014-01-25 16:00:00

σ (Real Data)	RMSE	MASE
3.66	5.1	0.3672

Table B.3: Baseline - Error for Uniques Percentage forecast from 2013-12-26 04:00:00 to 2014-01-25 16:00:00

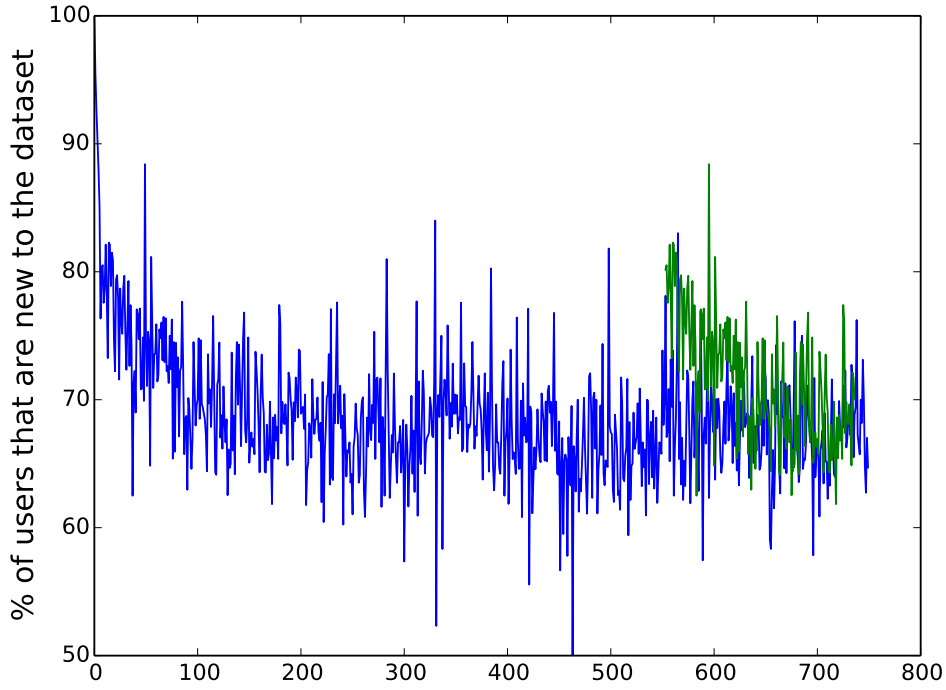

Figure B.4: Baseline - New uniques forecast from 2013-12-26 04:00:00 to 2014-01-25 16:00:00

σ (Real Data)	RMSE	MASE
3.61	6.69	0.4241

Table B.4: Baseline - Error for New Uniques forecast from 2013-12-26 04:00:00 to 2014-01-25 16:00:00

Case 2

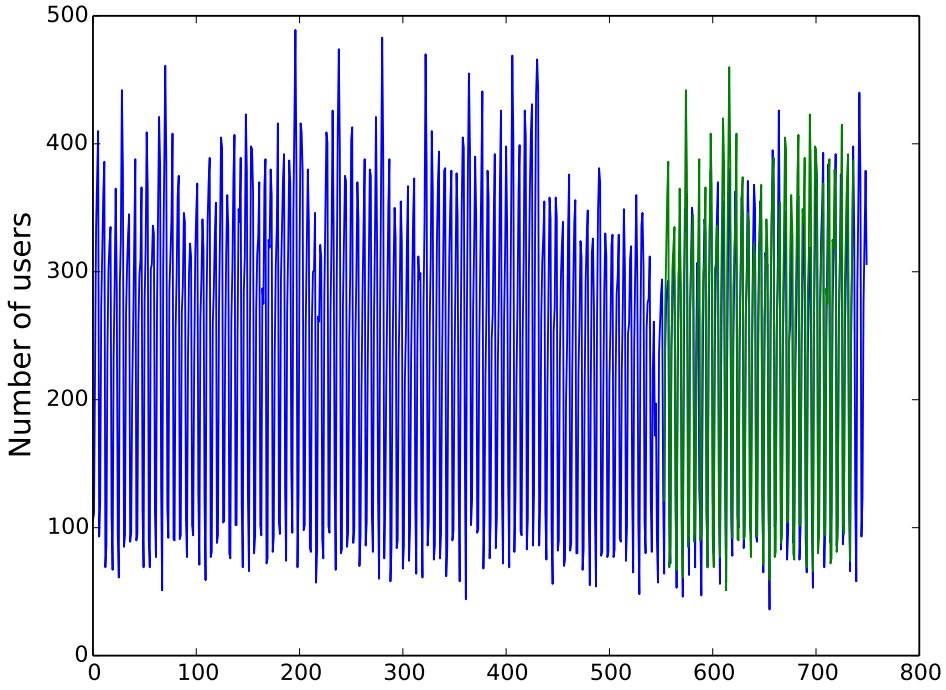

Figure B.5: Baseline - Uniques calculated using percentages forecast from 2013-12-26 04:00:00 to 2014-01-25 16:00:00

σ (Real Data)	RMSE	MASE
110.89	43.1	0.0992

Table B.5: Baseline - Error for Uniques calculated using percentages forecast from 2013-12-26 04:00:00 to 2014-01-25 16:00:00

B.2 Arima Allow Drift True - 4h

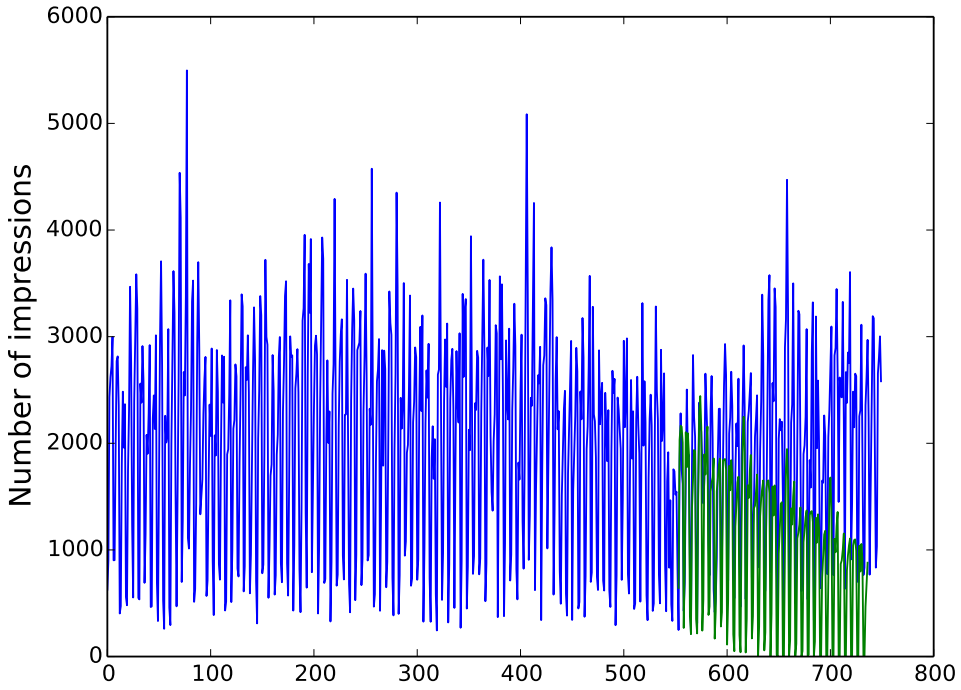

Figure B.6: Arima Allow Drift True - Impressions forecast from 2013-12-26 04:00:00 to 2014-01-25 16:00:00

σ (Real Data)	RMSE	MASE
908.7	1083.22	0.3152

Table B.6: Arima Allow Drift True - Error for Impressions forecast from 2013-12-26 04:00:00 to 2014-01-25 16:00:00

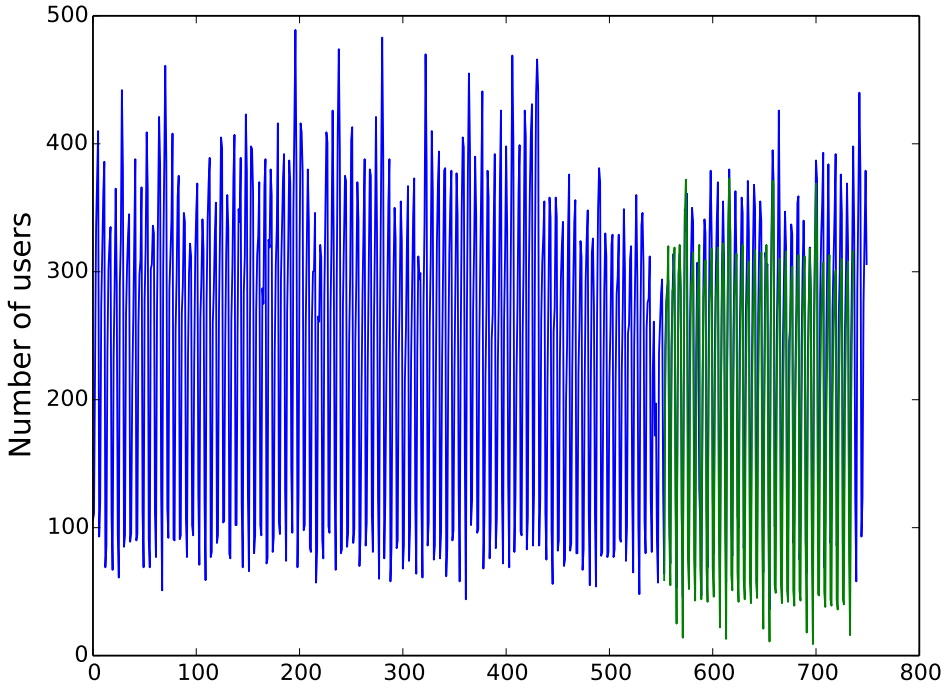

Figure B.7: Arima Allow Drift True - Uniques forecast from 2013-12-26 04:00:00 to 2014-01-25 16:00:00

σ (Real Data)	RMSE	MASE
110.89	46.24	0.1261

Table B.7: Arima Allow Drift True - Error for Uniques forecast from 2013-12-26 04:00:00 to 2014-01-25 16:00:00

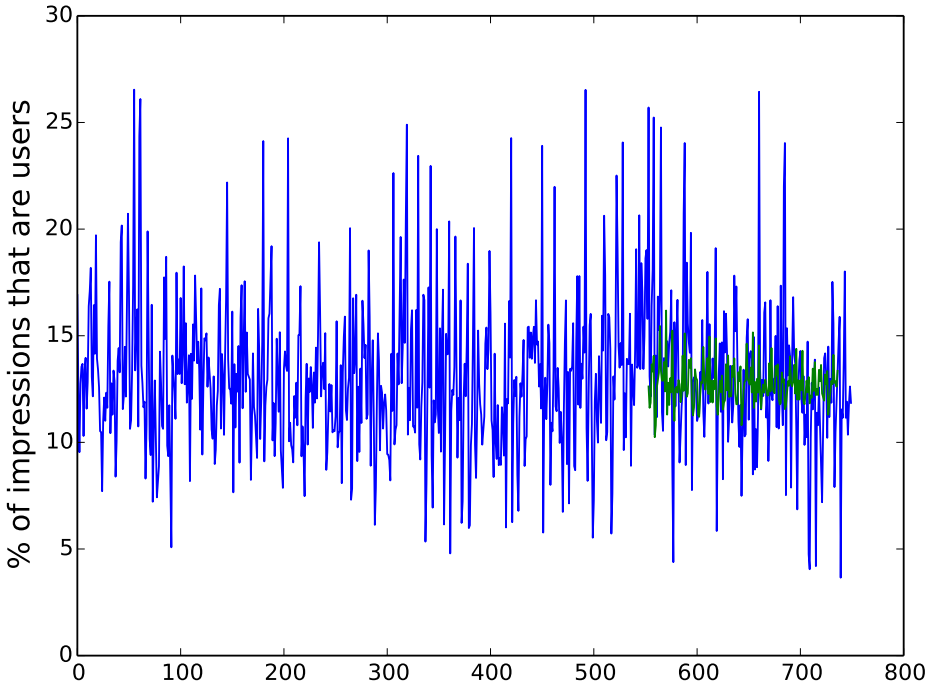

Figure B.8: Arima Allow Drift True - Uniques Percentage forecast from 2013-12-26 04:00:00 to 2014-01-25 16:00:00

σ (Real Data)	RMSE	MASE
3.66	3.6	0.2544

Table B.8: Arima Allow Drift True - Error for Uniques Percentage forecast from 2013-12-26 04:00:00 to 2014-01-25 16:00:00

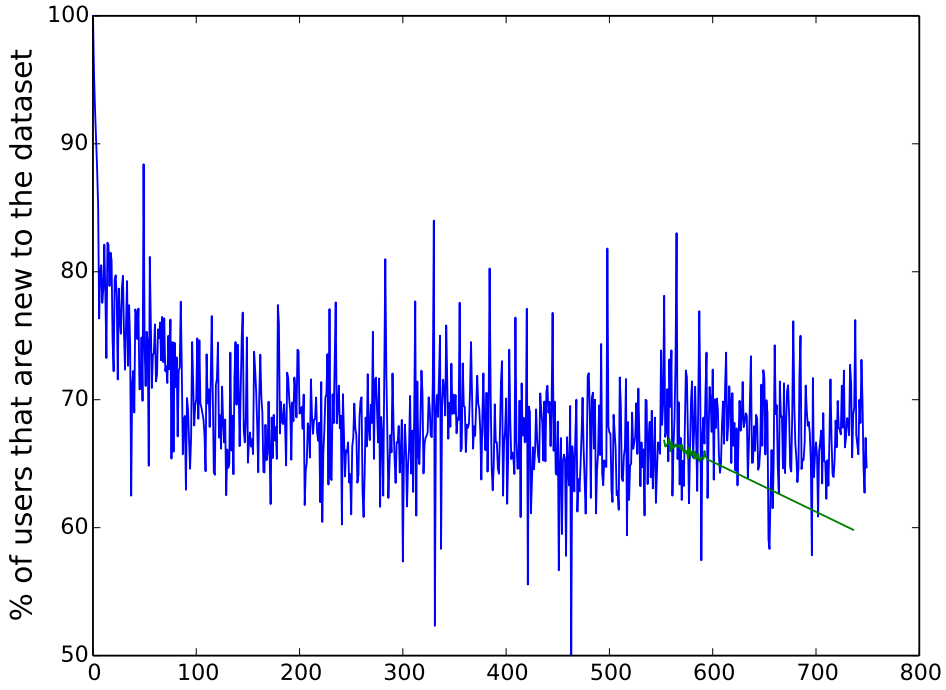

Figure B.9: Arima Allow Drift True - New uniques forecast from 2013-12-26 04:00:00 to 2014-01-25 16:00:00

σ (Real Data)	RMSE	MASE
3.61	5.79	0.3996

Table B.9: Arima Allow Drift True - Error for New Uniques forecast from 2013-12-26 04:00:00 to 2014-01-25 16:00:00

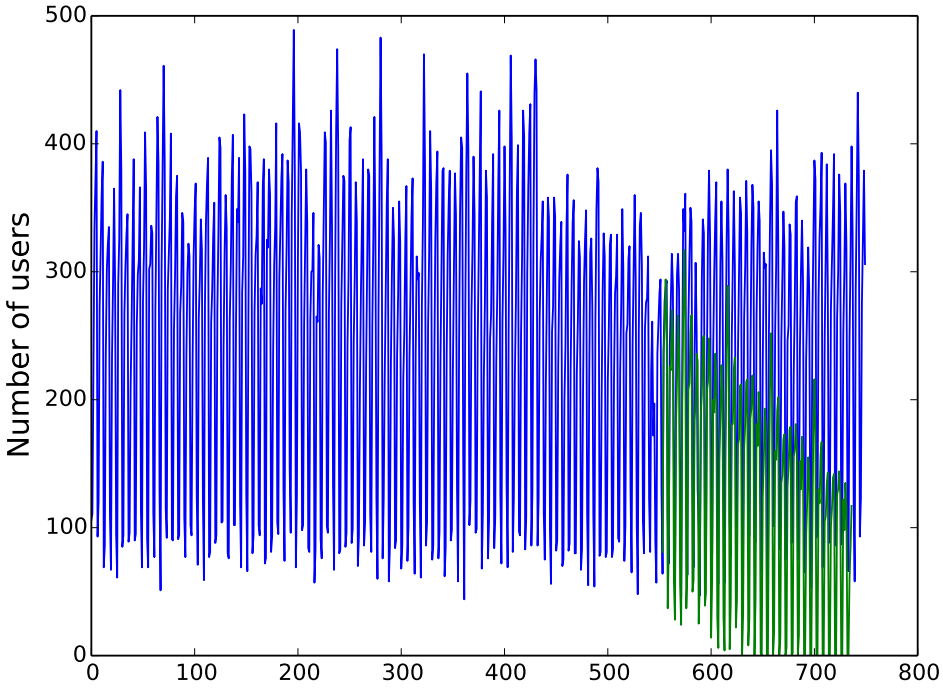

Figure B.10: Arima Allow Drift True - Uniques calculated using percentages forecast from 2013-12-26 04:00:00 to 2014-01-25 16:00:00

σ (Real Data)	RMSE	MASE
110.89	121.21	0.348

Table B.10: Arima Allow Drift True - Error for Uniques calculated using percentages forecast from 2013-12-26 04:00:00 to 2014-01-25 16:00:00

B.3 Arima Allow Drift False - 4h

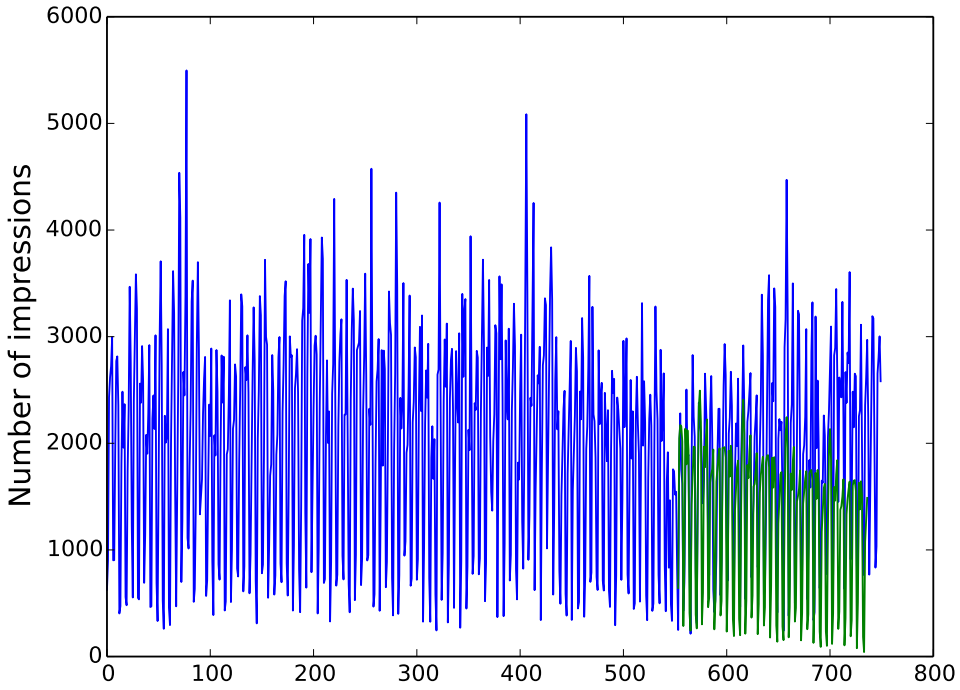

Figure B.11: Arima Allow Drift False - Impressions forecast from 2013-12-26 04:00:00 to 2014-01-25 16:00:00

σ (Real Data)	RMSE	MASE
908.7	834.32	0.2348

Table B.11: Arima Allow Drift False - Error for Impressions forecast from 2013-12-26 04:00:00 to 2014-01-25 16:00:00

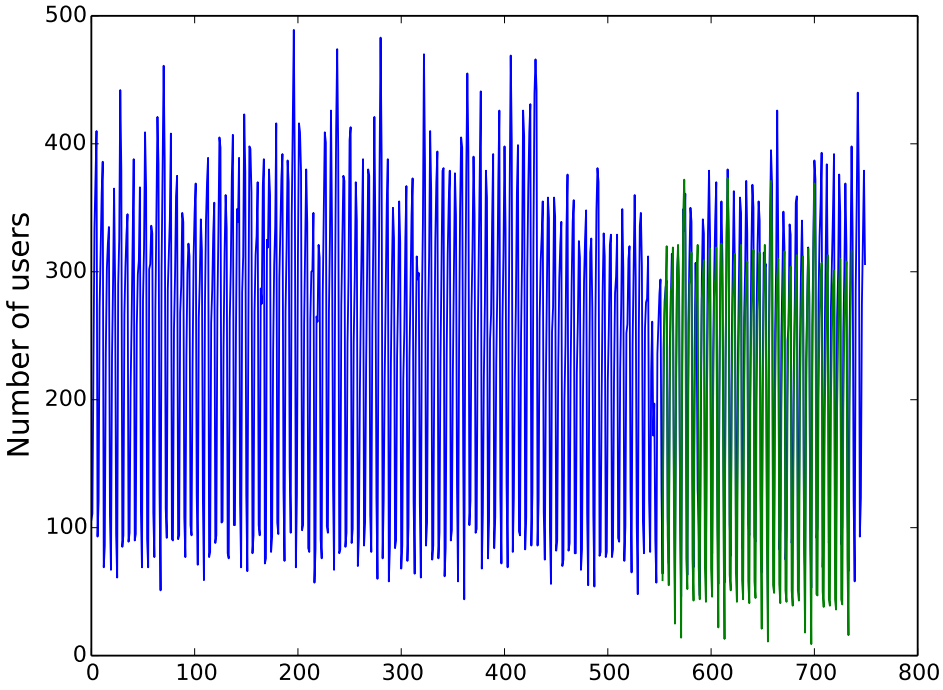

Figure B.12: Arima Allow Drift False - Uniques forecast from 2013-12-26 04:00:00 to 2014-01-25 16:00:00

σ (Real Data)	RMSE	MASE
110.89	46.24	0.1261

Table B.12: Arima Allow Drift False - Error for Uniques forecast from 2013-12-26 04:00:00 to 2014-01-25 16:00:00

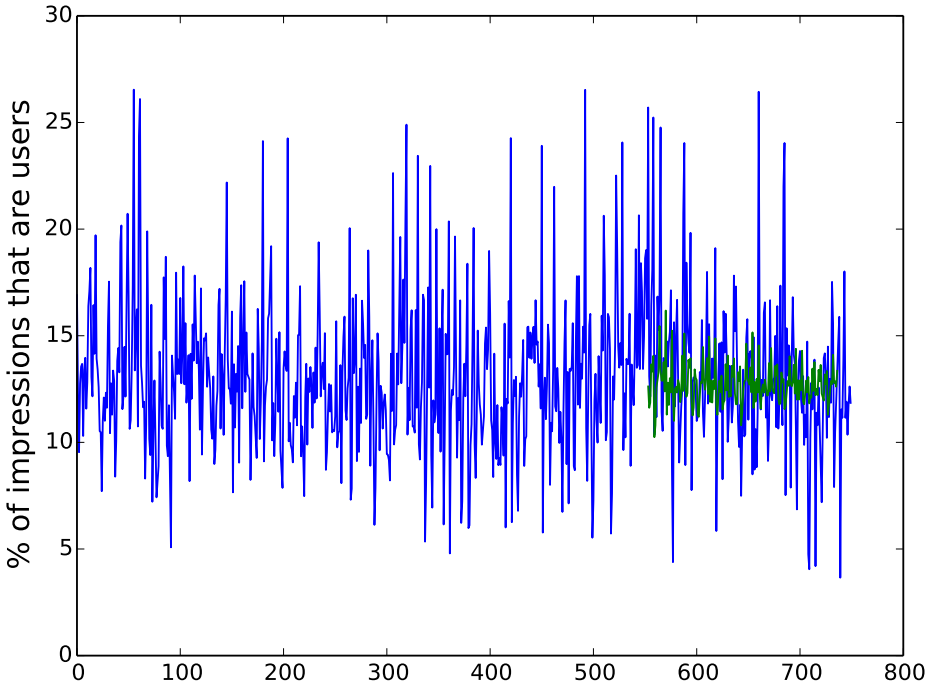

Figure B.13: Arima Allow Drift False - Uniques Percentage forecast from 2013-12-26 04:00:00 to 2014-01-25 16:00:00

σ (Real Data)	RMSE	MASE
3.66	3.6	0.2544

Table B.13: Arima Allow Drift False - Error for Uniques Percentage forecast from 2013-12-26 04:00:00 to 2014-01-25 16:00:00

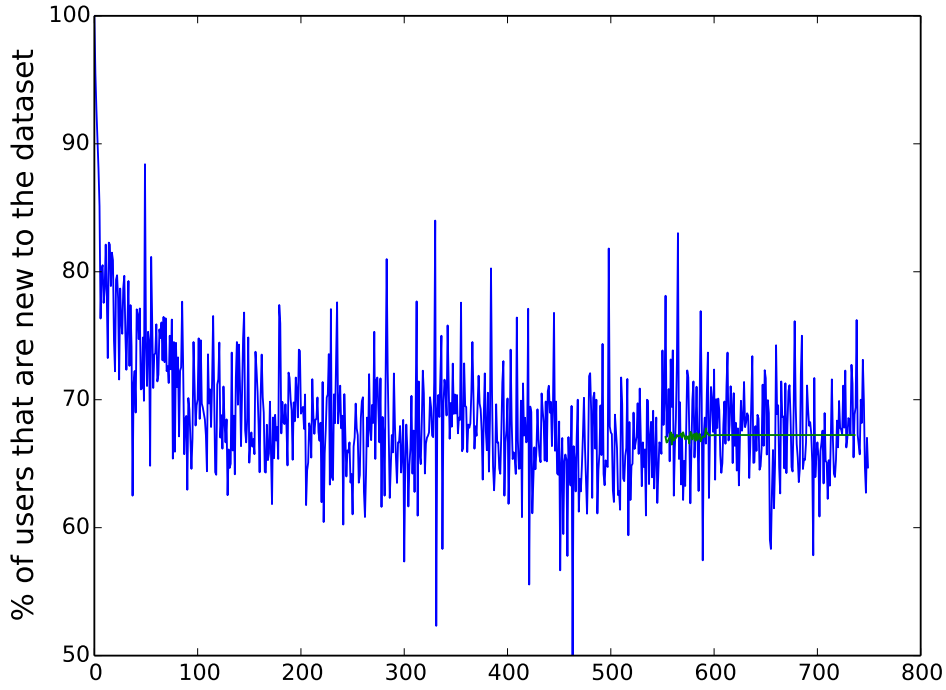

Figure B.14: Arima Allow Drift False - New uniques forecast from 2013-12-26 04:00:00 to 2014-01-25 16:00:00

σ (Real Data)	RMSE	MASE
3.61	3.66	0.2296

Table B.14: Arima Allow Drift False - Error for New Uniques forecast from 2013-12-26 04:00:00 to 2014-01-25 16:00:00

Case 2

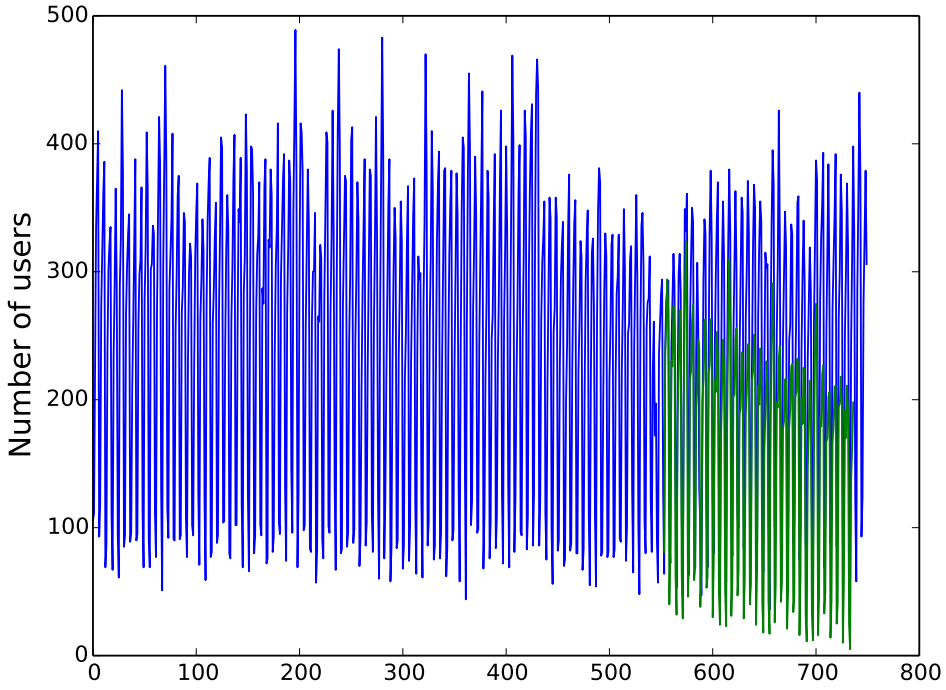

Figure B.15: Arima Allow Drift False - Uniques calculated using percentages forecast from 2013-12-26 04:00:00 to 2014-01-25 16:00:00

σ (Real Data)	RMSE	MASE
110.89	87.04	0.249

Table B.15: Arima Allow Drift False - Error for Uniques calculated using percentages forecast from 2013-12-26 04:00:00 to 2014-01-25 16:00:00

B.4 Baseline - 6h

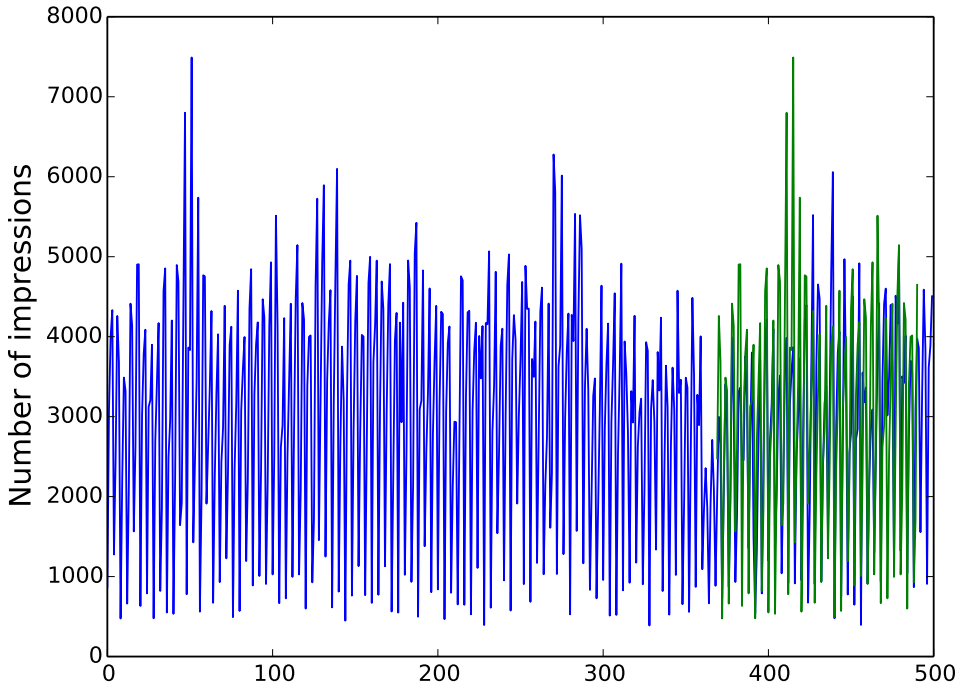

Figure B.16: Baseline - Impressions forecast from 2013-12-26 06:00:00 to 2014-01-25 12:00:00

σ (Real Data)	RMSE	MASE
1293.78	967.46	0.1312

Table B.16: Baseline - Error for Impressions forecast from 2013-12-26 06:00:00 to 2014-01-25 12:00:00

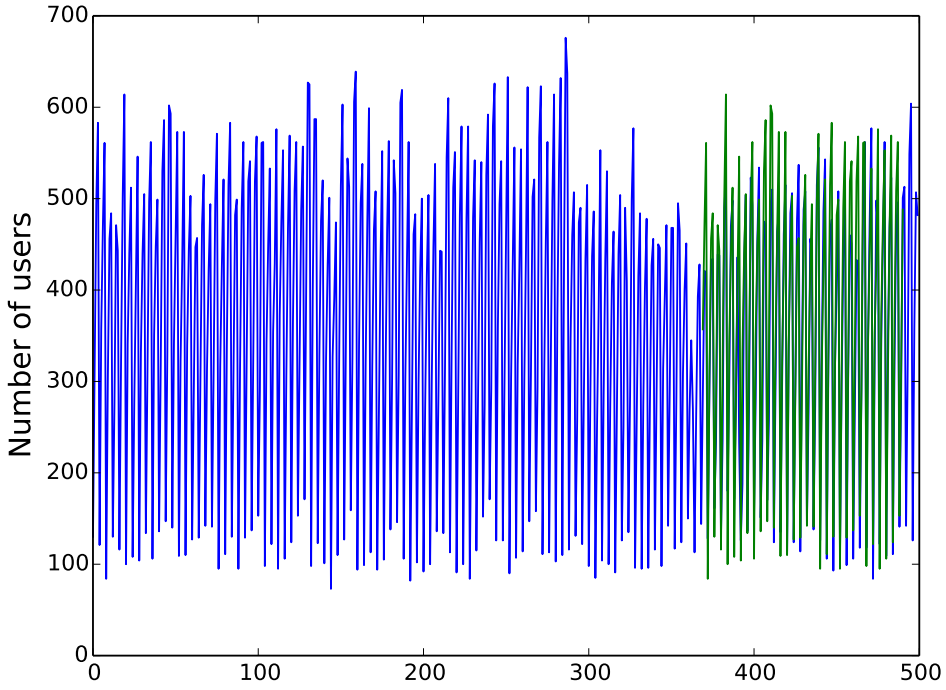

Figure B.17: Baseline - Uniques forecast from 2013-12-26 06:00:00 to 2014-01-25 12:00:00

σ (Real Data)	RMSE	MASE
152.99	57.47	0.063

Table B.17: Baseline - Error for Uniques forecast from 2013-12-26 06:00:00 to 2014-01-25 12:00:00

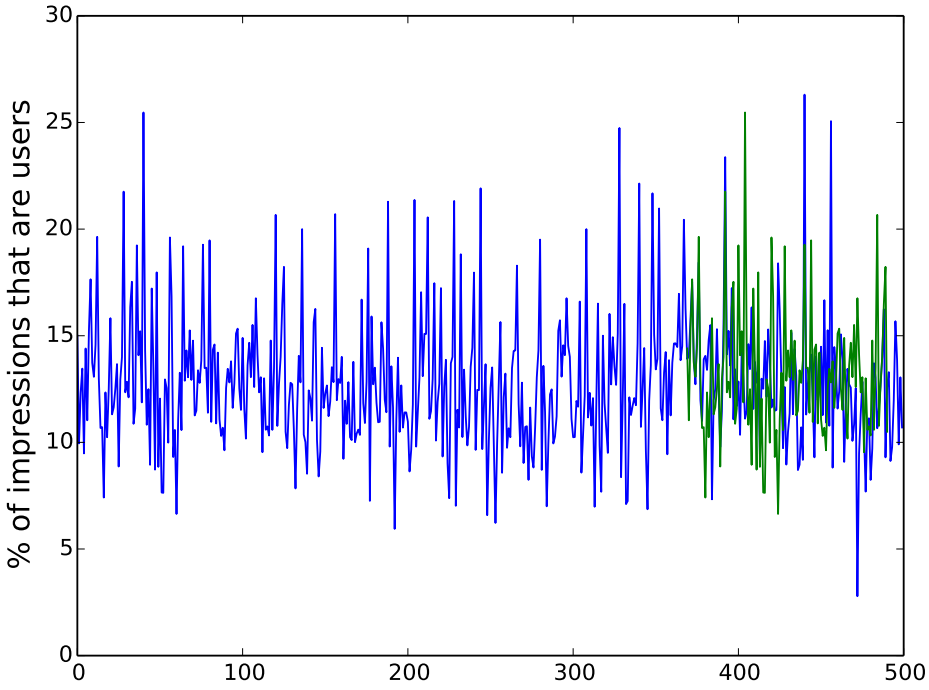

Figure B.18: Baseline - Uniques Percentage forecast from 2013-12-26 06:00:00 to 2014-01-25 12:00:00

σ (Real Data)	RMSE	MASE
3.01	4.15	0.2927

Table B.18: Baseline - Error for Uniques Percentage forecast from 2013-12-26 06:00:00 to 2014-01-25 12:00:00

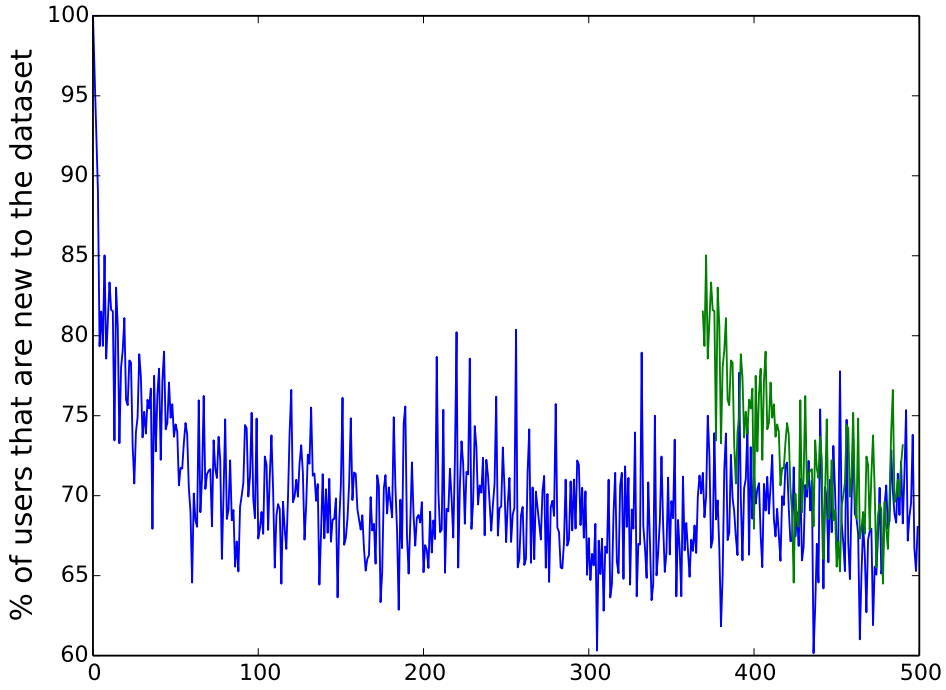

Figure B.19: Baseline - New uniques forecast from 2013-12-26 06:00:00 to 2014-01-25 12:00:00

σ (Real Data)	RMSE	MASE
3.14	6.14	0.4809

Table B.19: Baseline - Error for New Uniques forecast from 2013-12-26 06:00:00 to 2014-01-25 12:00:00

Case 2

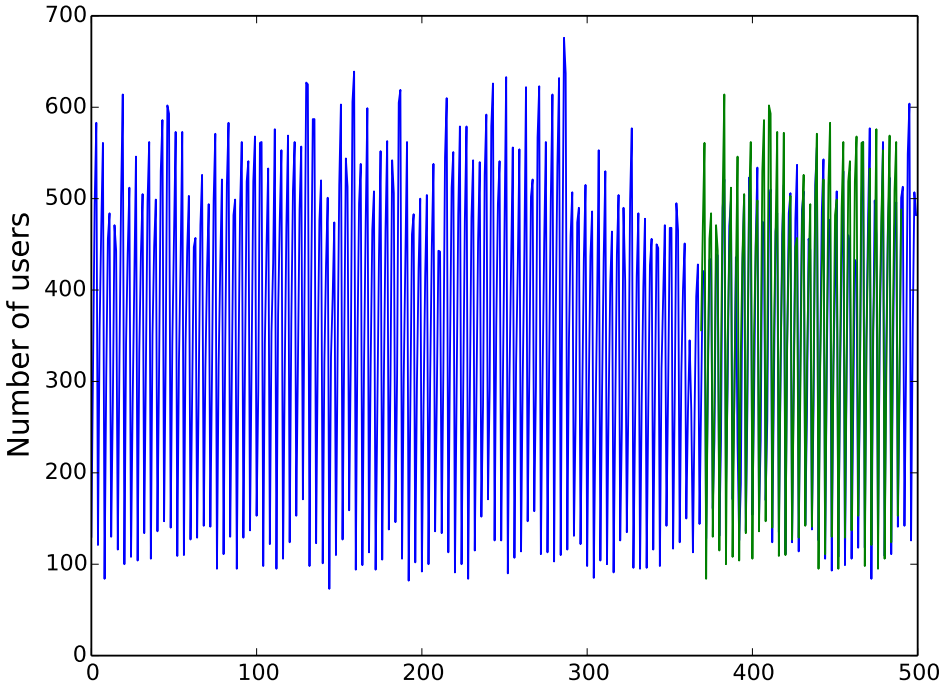

Figure B.20: Baseline - Uniques calculated using percentages forecast from 2013-12-26 06:00:00 to 2014-01-25 12:00:00

σ (Real Data)	RMSE	MASE
152.99	57.47	0.063

Table B.20: Baseline - Error for Uniques calculated using percentages forecast from 2013-12-26 06:00:00 to 2014-01-25 12:00:00

B.5 Arima Allow Drift True - 6h

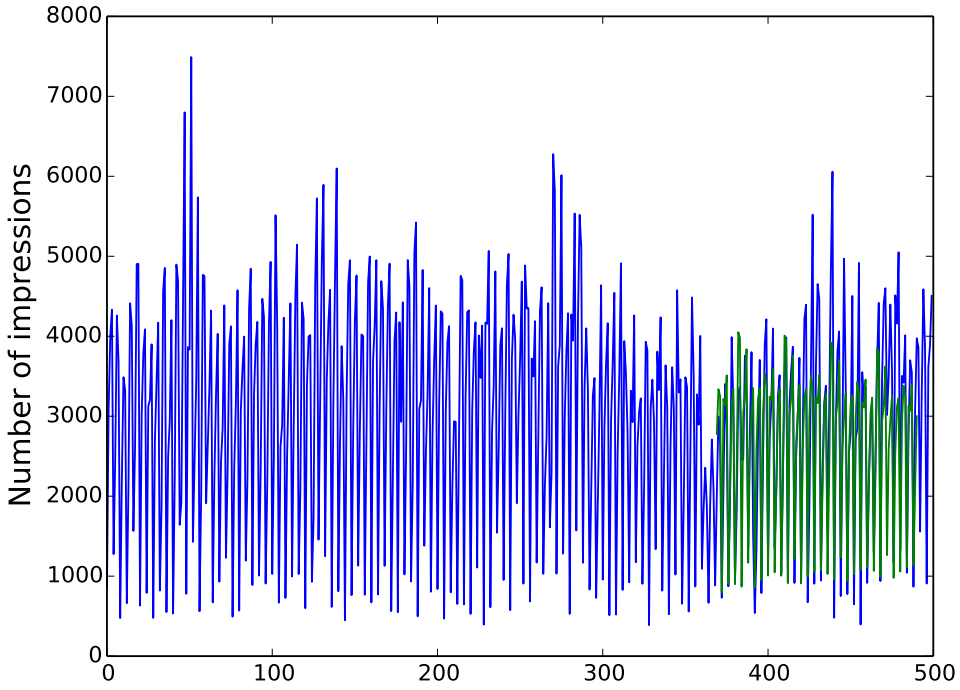

Figure B.21: Arima Allow Drift True - Impressions forecast from 2013-12-26 06:00:00 to 2014-01-25 12:00:00

σ (Real Data)	RMSE	MASE
1293.78	740.73	0.0947

Table B.21: Arima Allow Drift True - Error for Impressions forecast from 2013-12-26 06:00:00 to 2014-01-25 12:00:00

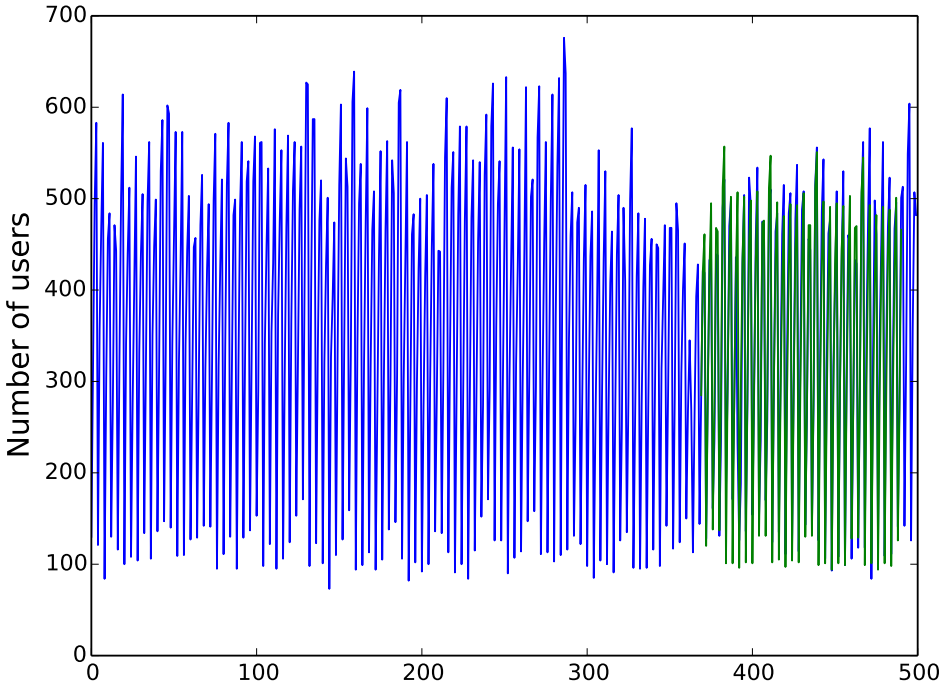

Figure B.22: Arima Allow Drift True - Uniques forecast from 2013-12-26 06:00:00 to 2014-01-25 12:00:00

σ (Real Data)	RMSE	MASE
152.99	42.43	0.0456

Table B.22: Arima Allow Drift True - Error for Uniques forecast from 2013-12-26 06:00:00 to 2014-01-25 12:00:00

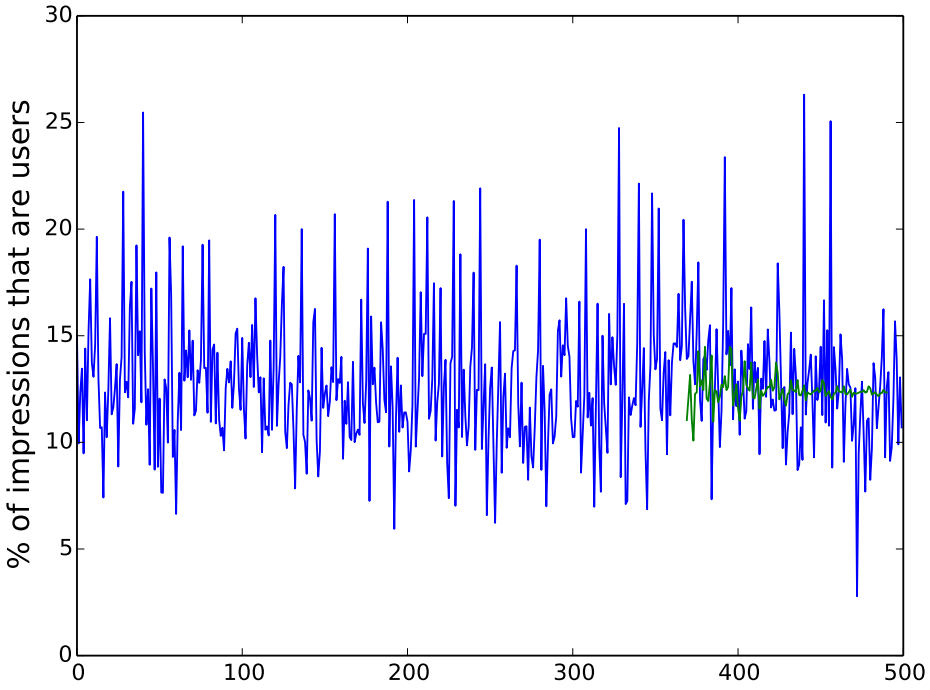

Figure B.23: Arima Allow Drift True - Uniques Percentage forecast from 2013-12-26 06:00:00 to 2014-01-25 12:00:00

σ (Real Data)	RMSE	MASE
3.01	3.05	0.2099

Table B.23: Arima Allow Drift True - Error for Uniques Percentage forecast from 2013-12-26 06:00:00 to 2014-01-25 12:00:00

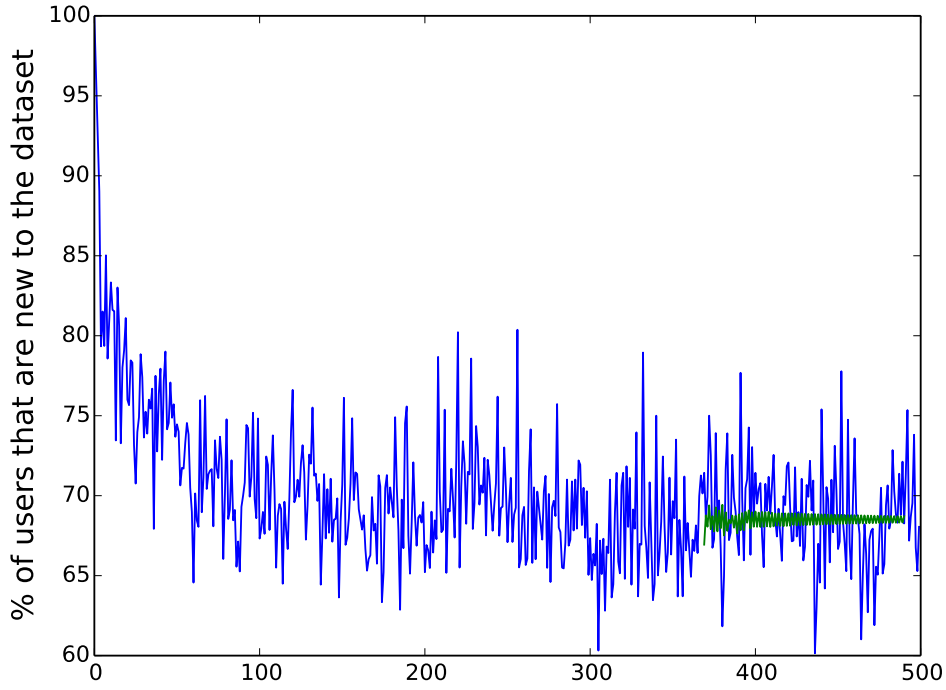

Figure B.24: Arima Allow Drift True - New uniques forecast from 2013-12-26 06:00:00 to 2014-01-25 12:00:00

σ (Real Data)	RMSE	MASE
3.14	3.16	0.2409

Table B.24: Arima Allow Drift True - Error for New Uniques forecast from 2013-12-26 06:00:00 to 2014-01-25 12:00:00

Case 2

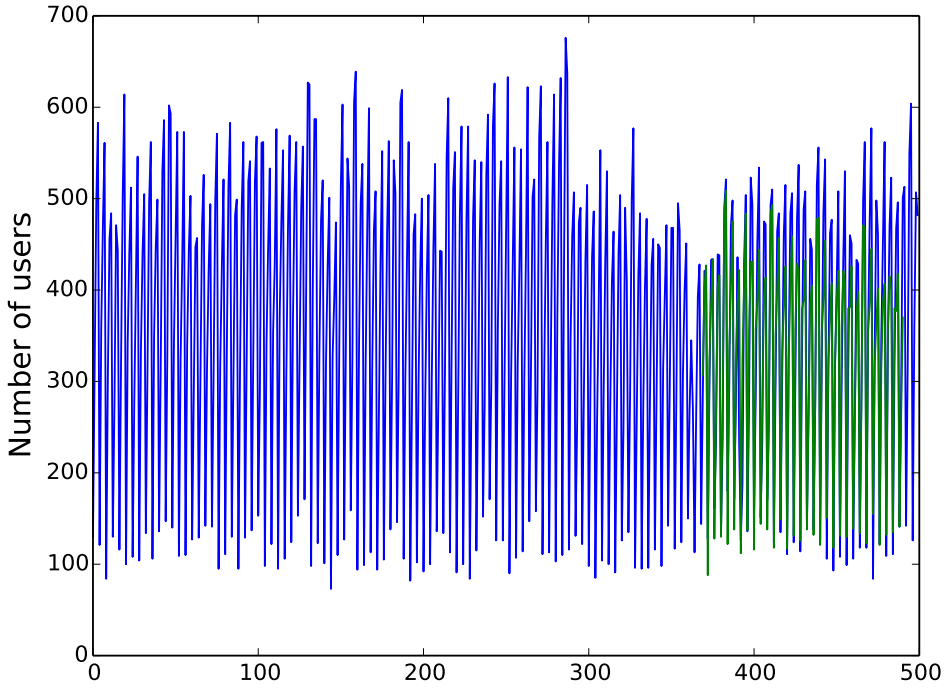

Figure B.25: Arima Allow Drift True - Uniques calculated using percentages forecast from 2013-12-26 06:00:00 to 2014-01-25 12:00:00

σ (Real Data)	RMSE	MASE
152.99	60.31	0.0724

Table B.25: Arima Allow Drift True - Error for Uniques calculated using percentages forecast from 2013-12-26 06:00:00 to 2014-01-25 12:00:00

B.6 Arima Allow Drift False - 6h

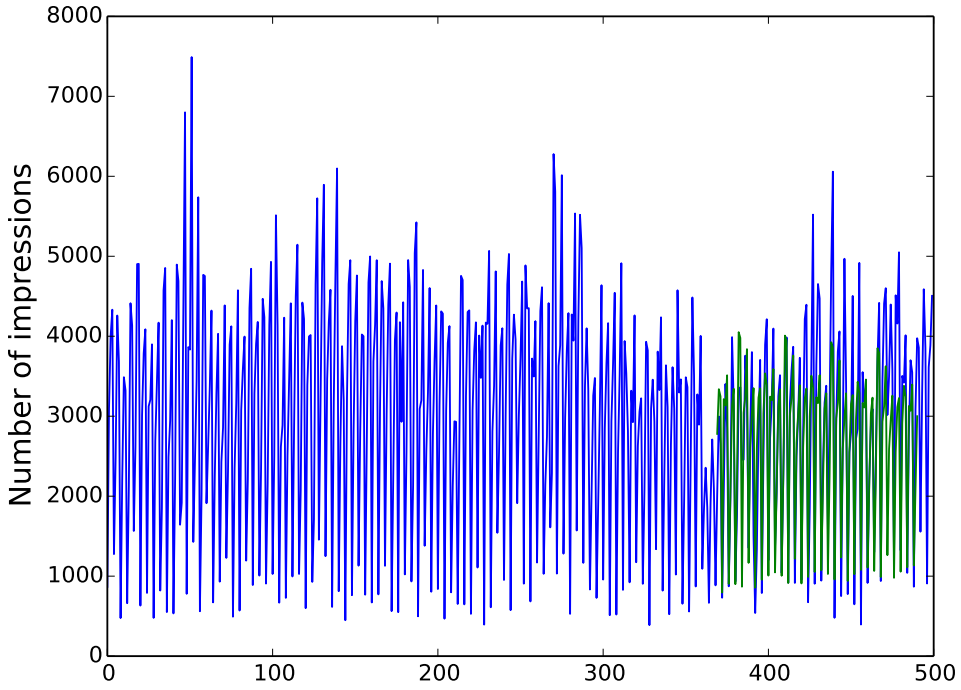

Figure B.26: Arima Allow Drift False - Impressions forecast from 2013-12-26 06:00:00 to 2014-01-25 12:00:00

σ (Real Data)	RMSE	MASE
1293.78	740.73	0.0947

Table B.26: Arima Allow Drift False - Error for Impressions forecast from 2013-12-26 06:00:00 to 2014-01-25 12:00:00

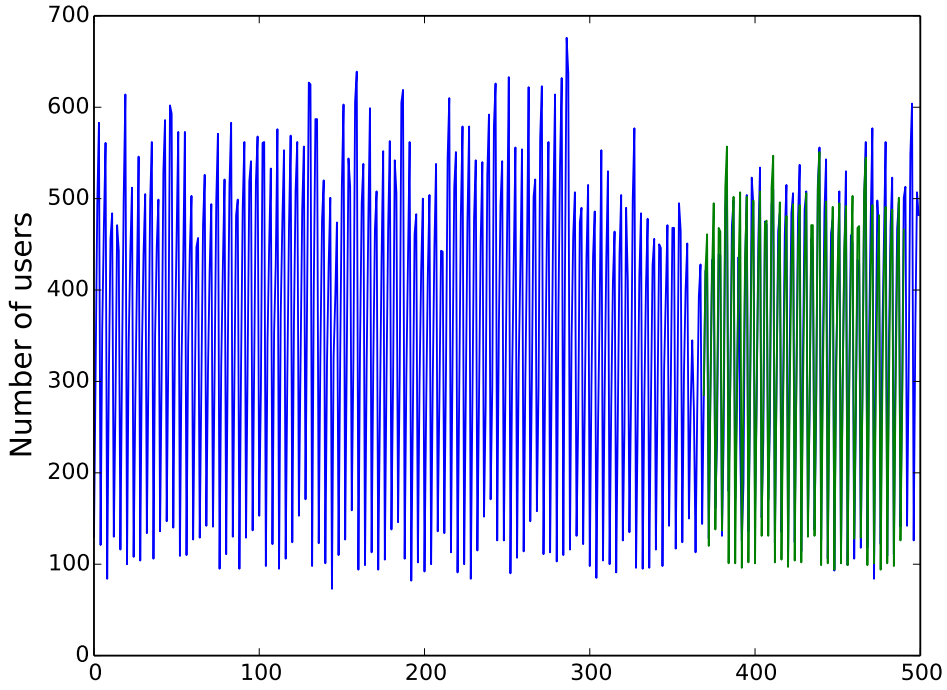

Figure B.27: Arima Allow Drift False - Uniques forecast from 2013-12-26 06:00:00 to 2014-01-25 12:00:00

σ (Real Data)	RMSE	MASE
152.99	42.43	0.0456

Table B.27: Arima Allow Drift False - Error for Uniques forecast from 2013-12-26 06:00:00 to 2014-01-25 12:00:00

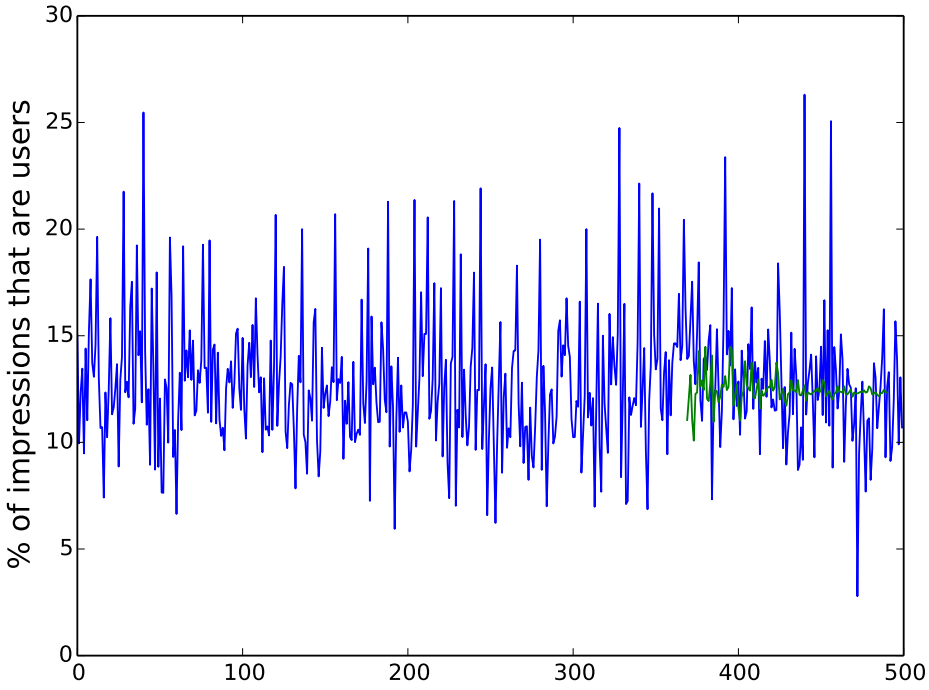

Figure B.28: Arima Allow Drift False - Uniques Percentage forecast from 2013-12-26 06:00:00 to 2014-01-25 12:00:00

σ (Real Data)	RMSE	MASE
3.01	3.05	0.2099

Table B.28: Arima Allow Drift False - Error for Uniques Percentage forecast from 2013-12-26 06:00:00 to 2014-01-25 12:00:00

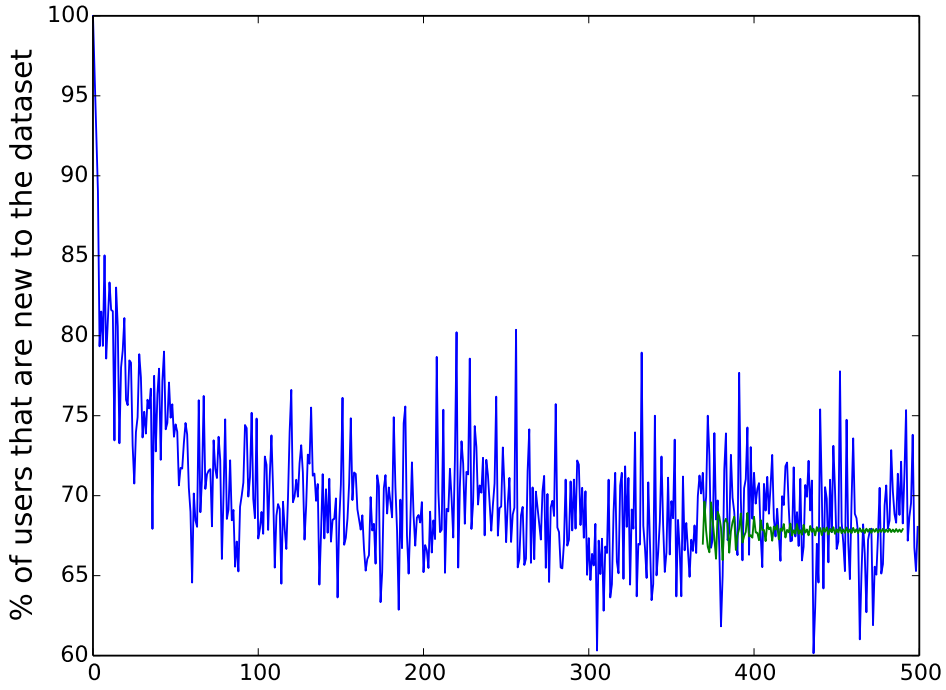

Figure B.29: Arima Allow Drift False - New uniques forecast from 2013-12-26 06:00:00 to 2014-01-25 12:00:00

σ (Real Data)	RMSE	MASE
3.14	3.33	0.2577

Table B.29: Arima Allow Drift False - Error for New Uniques forecast from 2013-12-26 06:00:00 to 2014-01-25 12:00:00

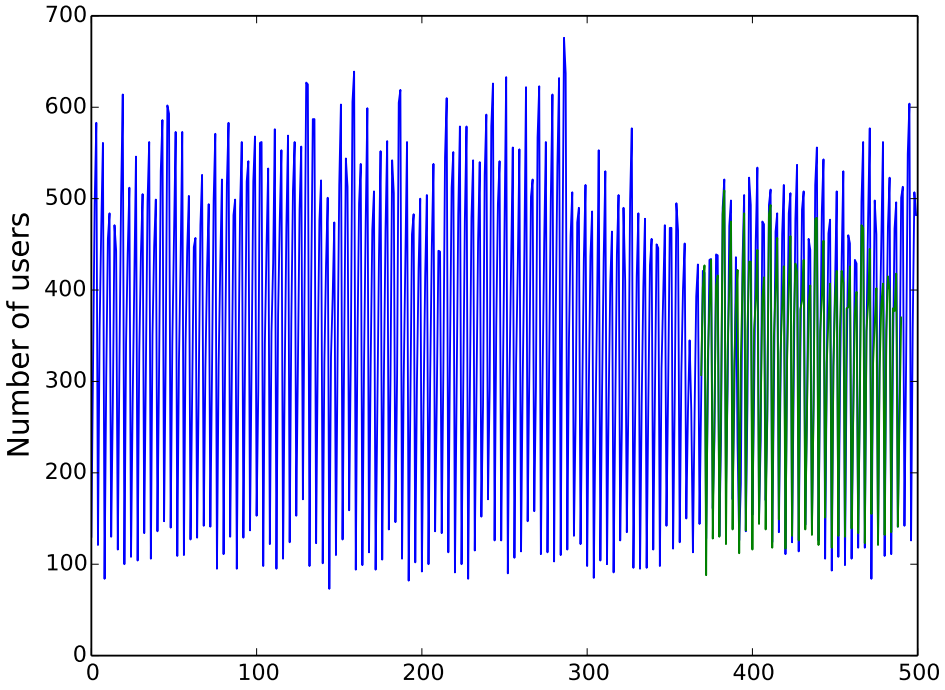

Figure B.30: Arima Allow Drift False - Uniques calculated using percentages forecast from 2013-12-26 06:00:00 to 2014-01-25 12:00:00

σ (Real Data)	RMSE	MASE
152.99	60.31	0.0724

Table B.30: Arima Allow Drift False - Error for Uniques calculated using percentages forecast from 2013-12-26 06:00:00 to 2014-01-25 12:00:00

B.7 Baseline - 8h

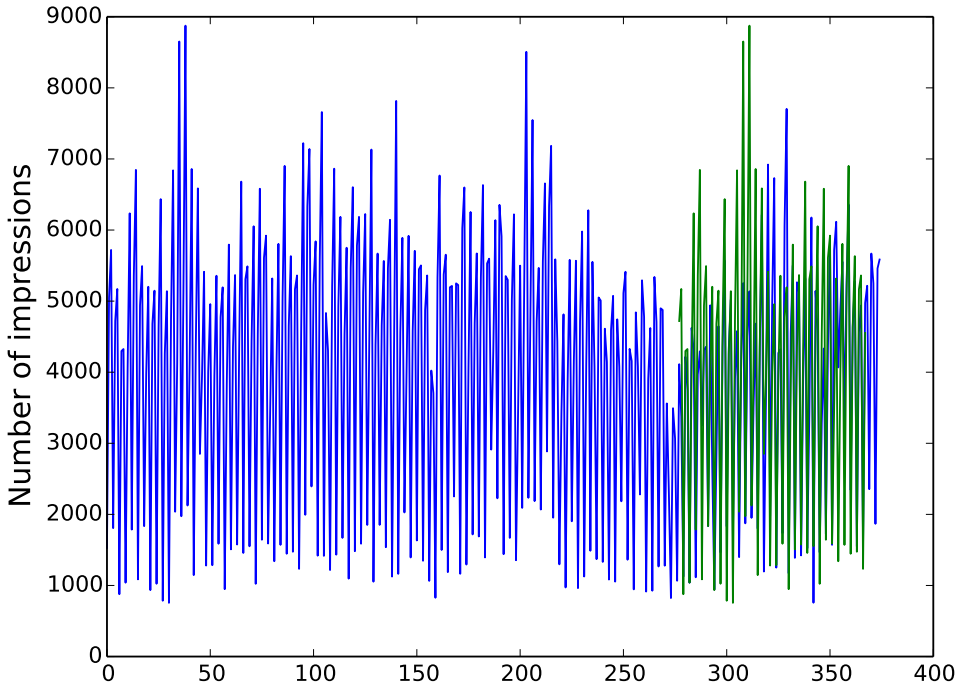

Figure B.31: Baseline - Impressions forecast from 2013-12-26 08:00:00 to 2014-01-25 08:00:00

σ (Real Data)	RMSE	MASE
1691.32	1205.49	0.0983

Table B.31: Baseline - Error for Impressions forecast from 2013-12-26 08:00:00 to 2014-01-25 08:00:00

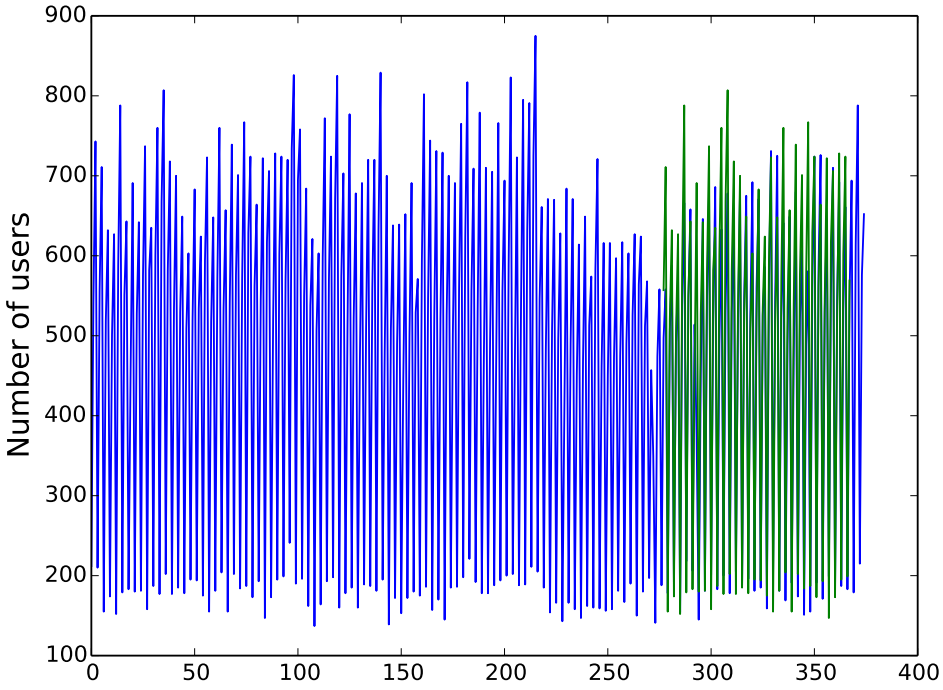

Figure B.32: Baseline - Uniques forecast from 2013-12-26 08:00:00 to 2014-01-25 08:00:00

σ (Real Data)	RMSE	MASE
203.33	66.02	0.0438

Table B.32: Baseline - Error for Uniques forecast from 2013-12-26 08:00:00 to 2014-01-25 08:00:00

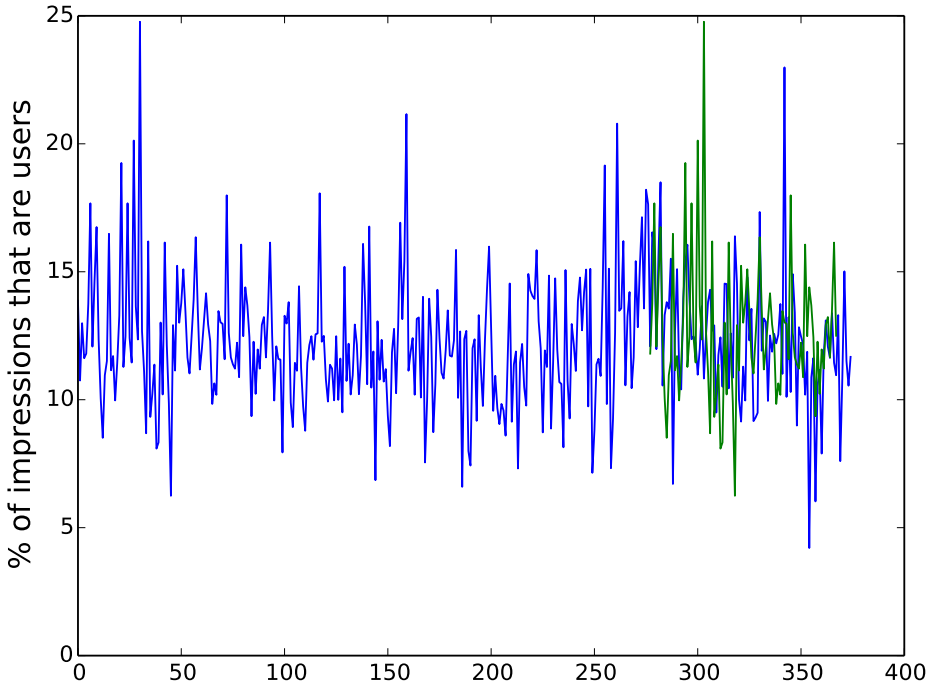

Figure B.33: Baseline - Uniques Percentage forecast from 2013-12-26 08:00:00 to 2014-01-25 08:00:00

σ (Real Data)	RMSE	MASE
2.57	3.82	0.336

Table B.33: Baseline - Error for Uniques Percentage forecast from 2013-12-26 08:00:00 to 2014-01-25 08:00:00

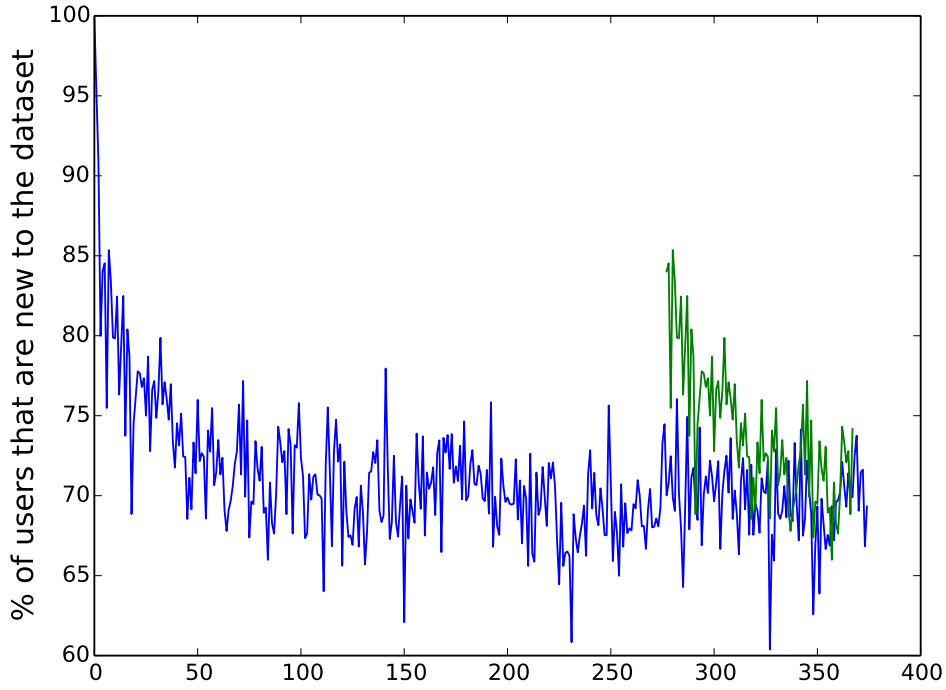

Figure B.34: Baseline - New uniques forecast from 2013-12-26 08:00:00 to 2014-01-25 08:00:00

σ (Real Data)	RMSE	MASE
2.5	5.97	0.5684

Table B.34: Baseline - Error for New Uniques forecast from 2013-12-26 08:00:00 to 2014-01-25 08:00:00

Case 2

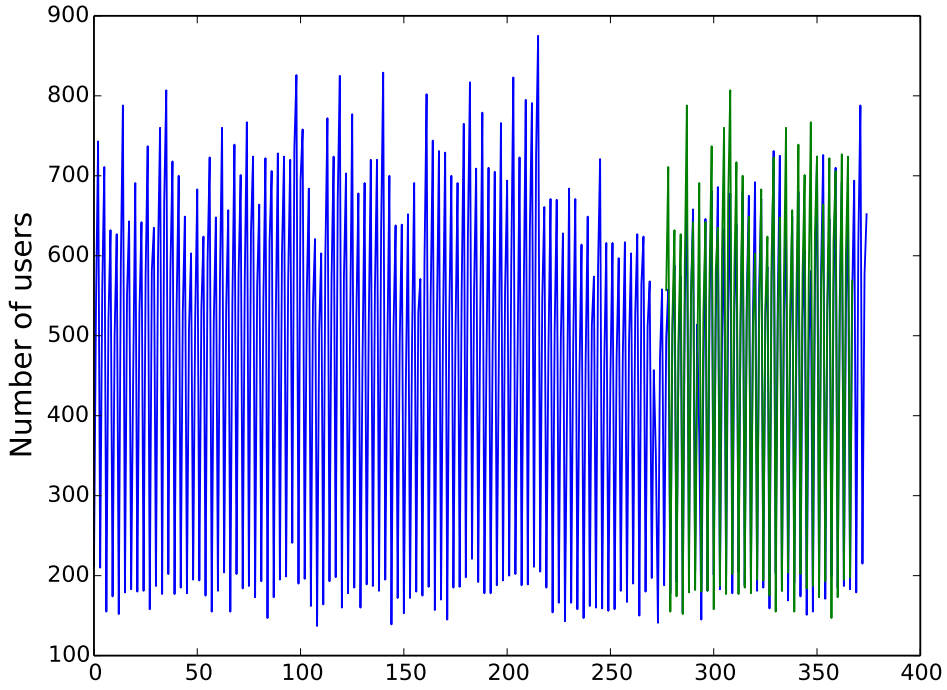

Figure B.35: Baseline - Uniques calculated using percentages forecast from 2013-12-26 08:00:00 to 2014-01-25 08:00:00

σ (Real Data)	RMSE	MASE
203.33	65.99	0.0437

Table B.35: Baseline - Error for Uniques calculated using percentages forecast from 2013-12-26 08:00:00 to 2014-01-25 08:00:00

B.8 Arima Allow Drift True - 8h

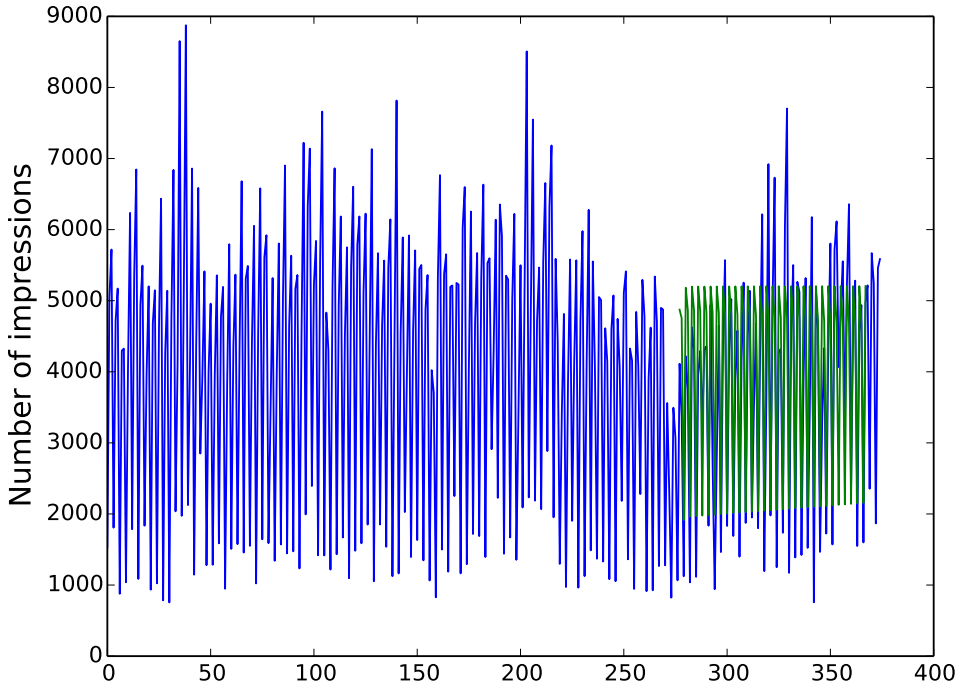

Figure B.36: Arima Allow Drift True - Impressions forecast from 2013-12-26 08:00:00 to 2014-01-25 08:00:00

σ (Real Data)	RMSE	MASE
1691.32	999.47	0.0905

Table B.36: Arima Allow Drift True - Error for Impressions forecast from 2013-12-26 08:00:00 to 2014-01-25 08:00:00

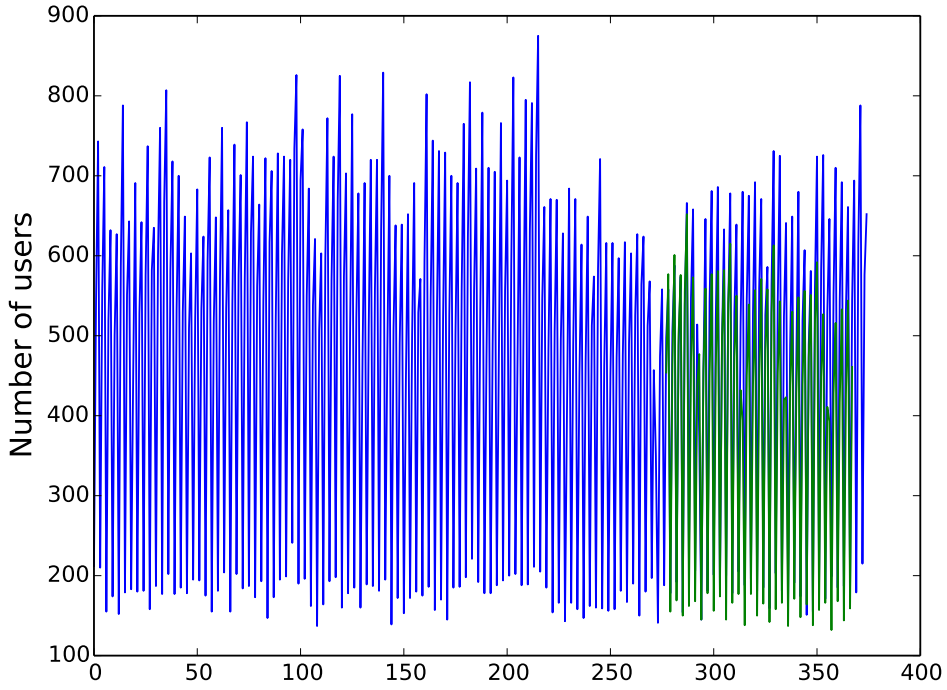

Figure B.37: Arima Allow Drift True - Uniques forecast from 2013-12-26 08:00:00 to 2014-01-25 08:00:00

σ (Real Data)	RMSE	MASE
203.33	89.79	0.0644

Table B.37: Arima Allow Drift True - Error for Uniques forecast from 2013-12-26 08:00:00 to 2014-01-25 08:00:00

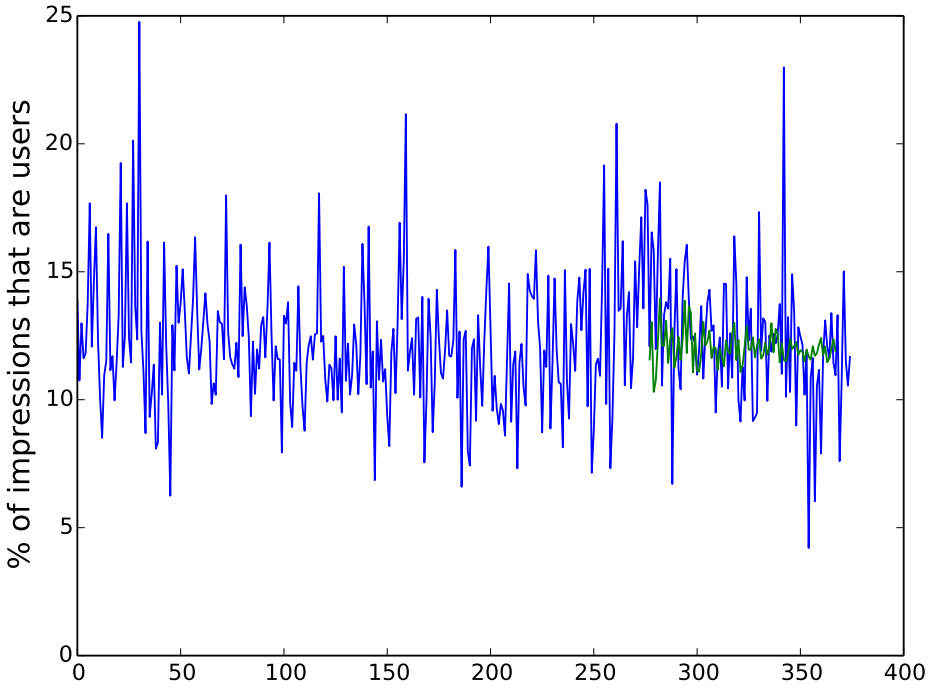

Figure B.38: Arima Allow Drift True - Uniques Percentage forecast from 2013-12-26 08:00:00 to 2014-01-25 08:00:00

σ (Real Data)	RMSE	MASE
2.57	2.58	0.2275

Table B.38: Arima Allow Drift True - Error for Uniques Percentage forecast from 2013-12-26 08:00:00 to 2014-01-25 08:00:00

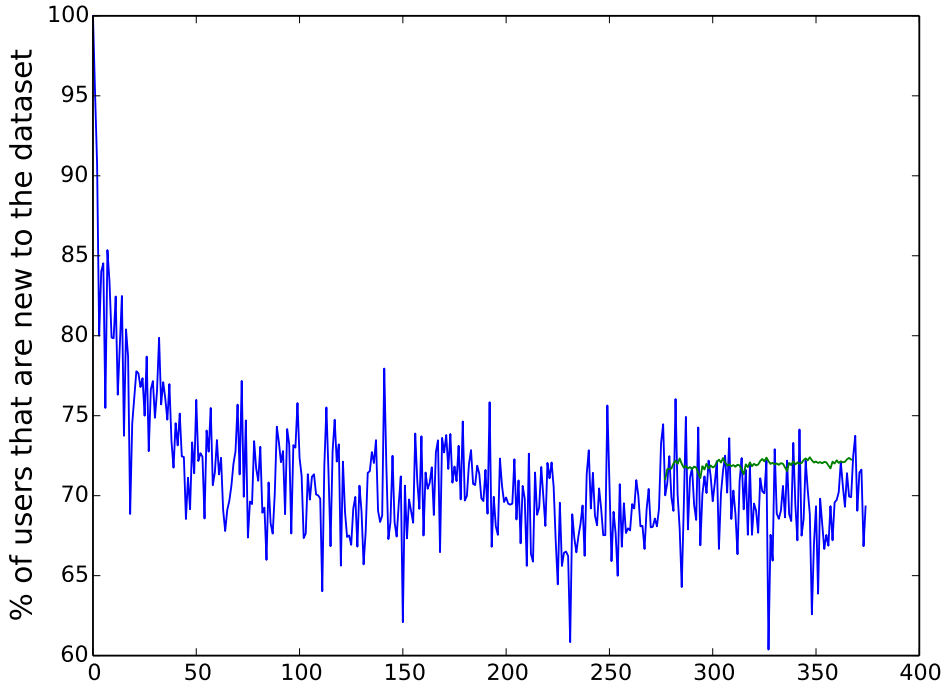

Figure B.39: Arima Allow Drift True - New uniques forecast from 2013-12-26 08:00:00 to 2014-01-25 08:00:00

σ (Real Data)	RMSE	MASE
2.5	3.45	0.3254

Table B.39: Arima Allow Drift True - Error for New Uniques forecast from 2013-12-26 08:00:00 to 2014-01-25 08:00:00

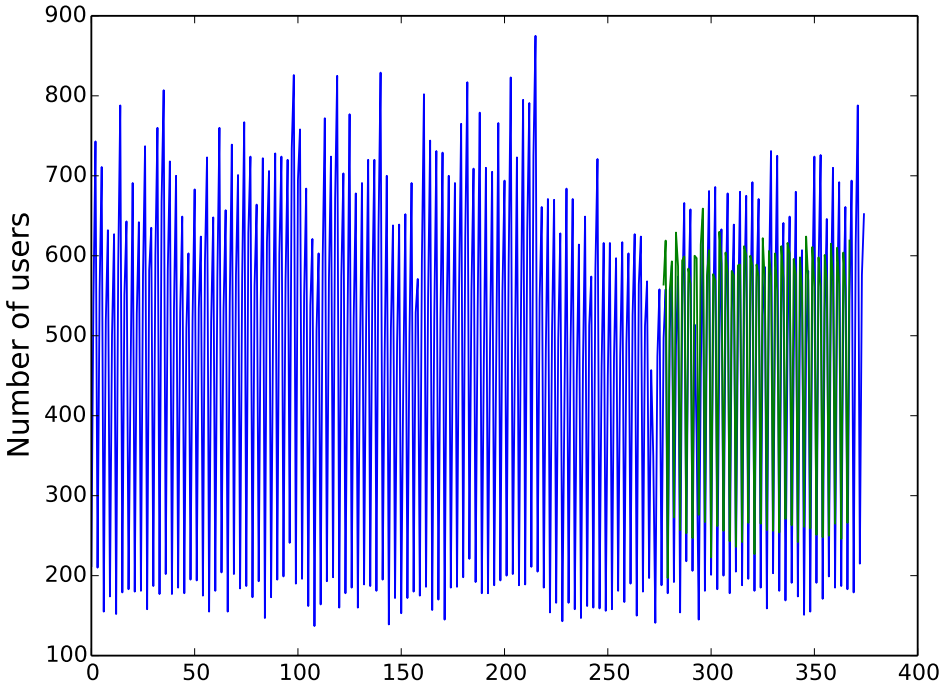

Figure B.40: Arima Allow Drift True - Uniques calculated using percentages forecast from 2013-12-26 08:00:00 to 2014-01-25 08:00:00

σ (Real Data)	RMSE	MASE
203.33	87.91	0.0738

Table B.40: Arima Allow Drift True - Error for Uniques calculated using percentages forecast from 2013-12-26 08:00:00 to 2014-01-25 08:00:00

B.9 Arima Allow Drift False - 8h

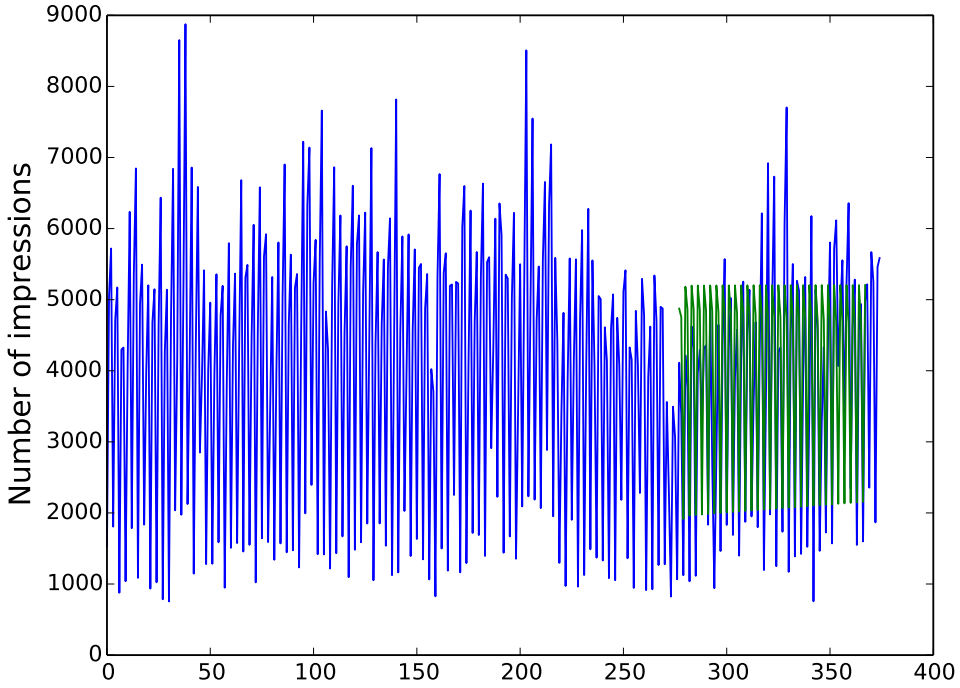

Figure B.41: Arima Allow Drift False - Impressions forecast from 2013-12-26 08:00:00 to 2014-01-25 08:00:00

σ (Real Data)	RMSE	MASE
1691.32	999.47	0.0905

Table B.41: Arima Allow Drift False - Error for Impressions forecast from 2013-12-26 08:00:00 to 2014-01-25 08:00:00

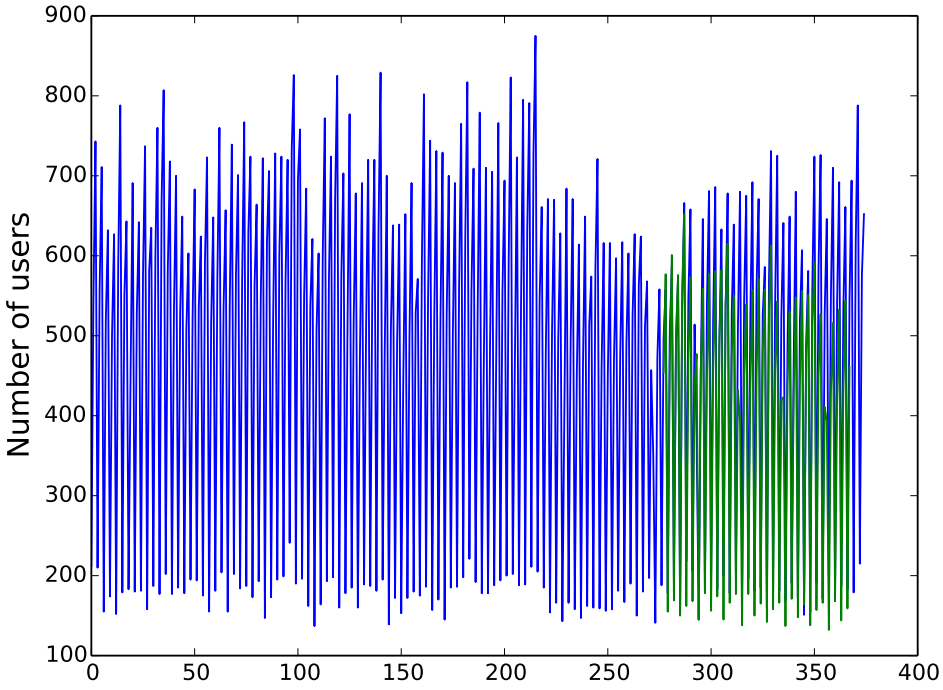

Figure B.42: Arima Allow Drift False - Uniques forecast from 2013-12-26 08:00:00 to 2014-01-25 08:00:00

σ (Real Data)	RMSE	MASE
203.33	89.79	0.0644

Table B.42: Arima Allow Drift False - Error for Uniques forecast from 2013-12-26 08:00:00 to 2014-01-25 08:00:00

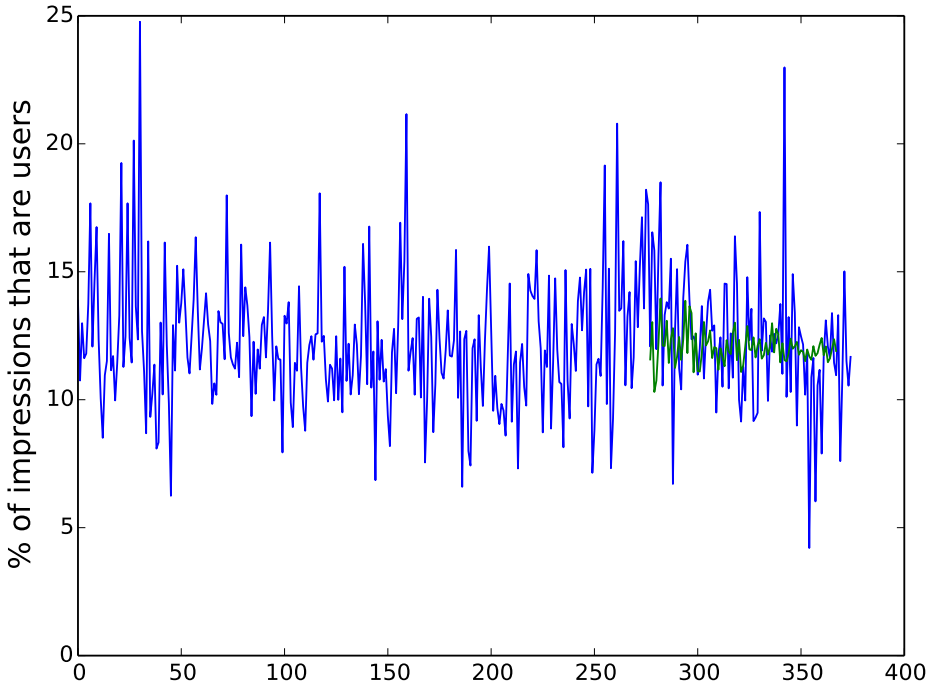

Figure B.43: Arima Allow Drift False - Uniques Percentage forecast from 2013-12-26 08:00:00 to 2014-01-25 08:00:00

σ (Real Data)	RMSE	MASE
2.57	2.58	0.2275

Table B.43: Arima Allow Drift False - Error for Uniques Percentage forecast from 2013-12-26 08:00:00 to 2014-01-25 08:00:00

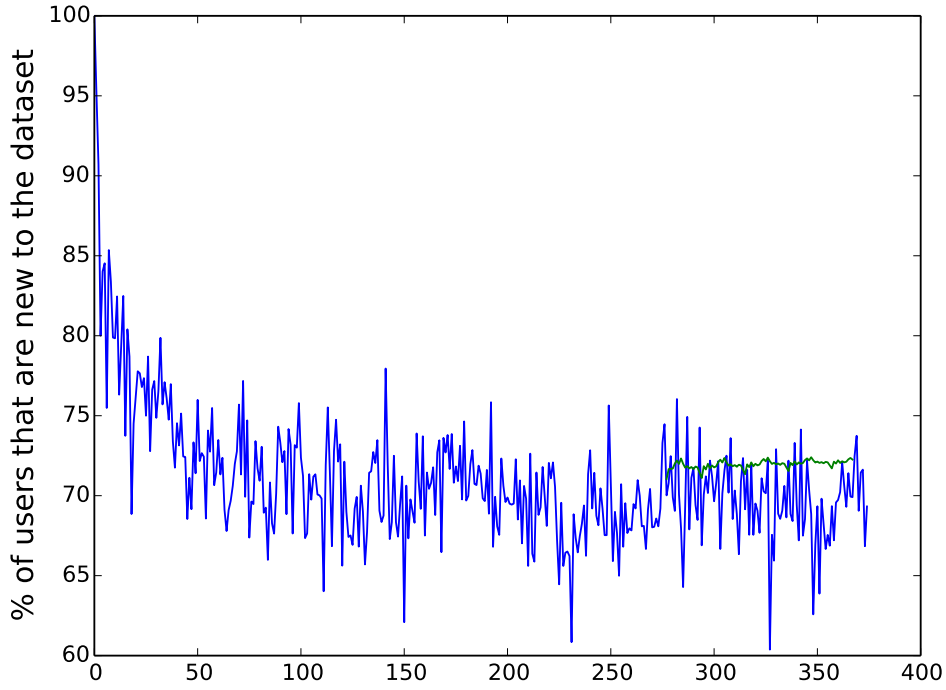

Figure B.44: Arima Allow Drift False - New uniques forecast from 2013-12-26 08:00:00 to 2014-01-25 08:00:00

σ (Real Data)	RMSE	MASE
2.5	3.45	0.3254

Table B.44: Arima Allow Drift False - Error for New Uniques forecast from 2013-12-26 08:00:00 to 2014-01-25 08:00:00

Case 2

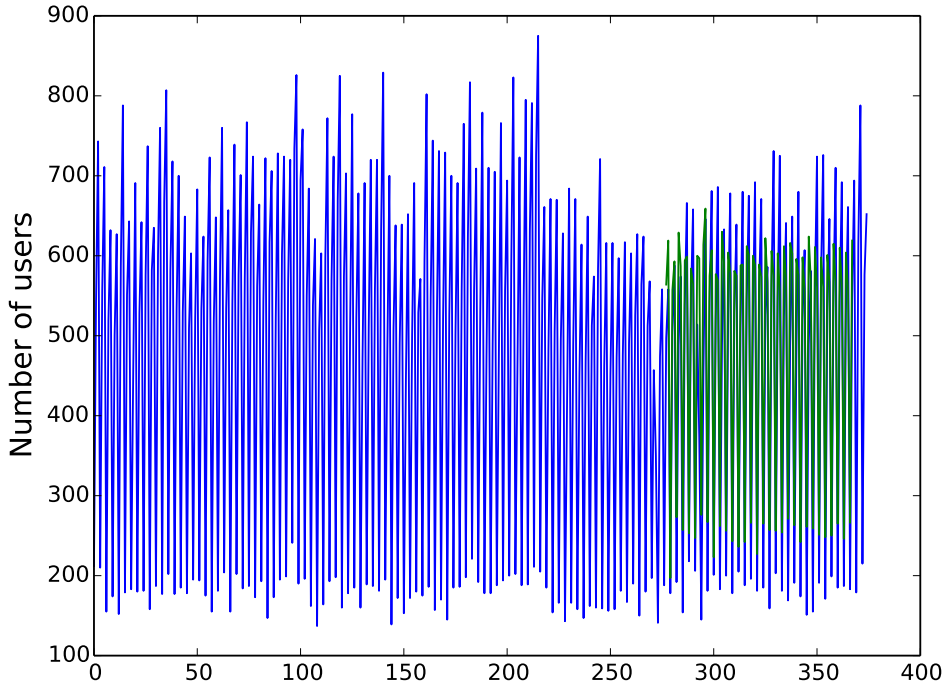

Figure B.45: Arima Allow Drift False - Uniques calculated using percentages forecast from 2013-12-26 08:00:00 to 2014-01-25 08:00:00

σ (Real Data)	RMSE	MASE
203.33	87.91	0.0738

Table B.45: Arima Allow Drift False - Error for Uniques calculated using percentages forecast from 2013-12-26 08:00:00 to 2014-01-25 08:00:00

B.10 Baseline - 12h

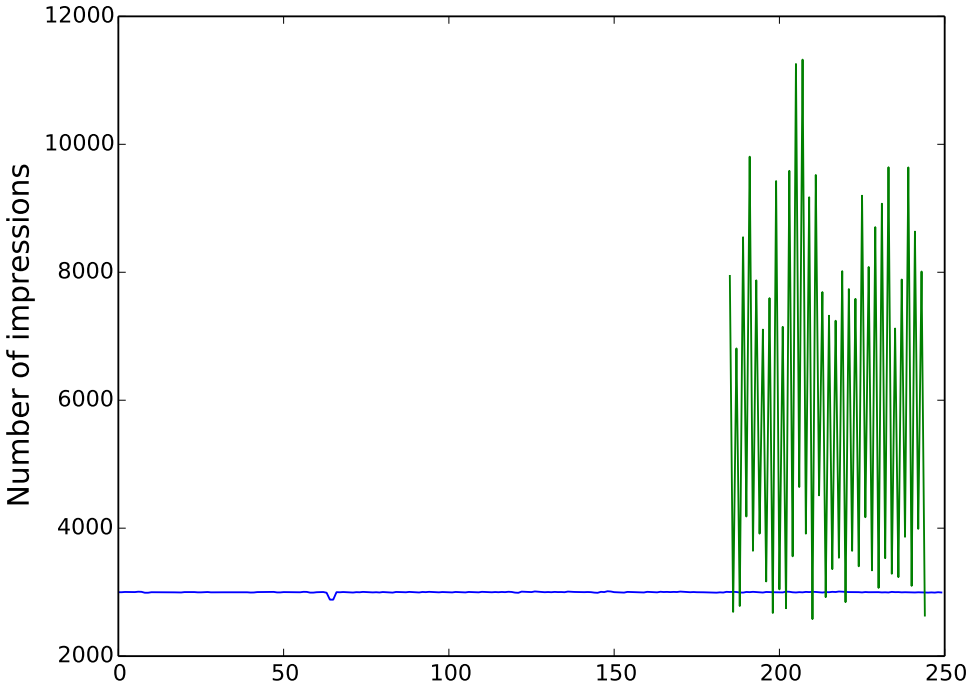

Figure B.46: Baseline - Impressions forecast from 2013-12-26 12:00:00 to 2014-01-25 00:00:00

σ (Real Data)	RMSE	MASE
3.79	3998.48	209.8582

Table B.46: Baseline - Error for Impressions forecast from 2013-12-26 12:00:00 to 2014-01-25 00:00:00

Case 2

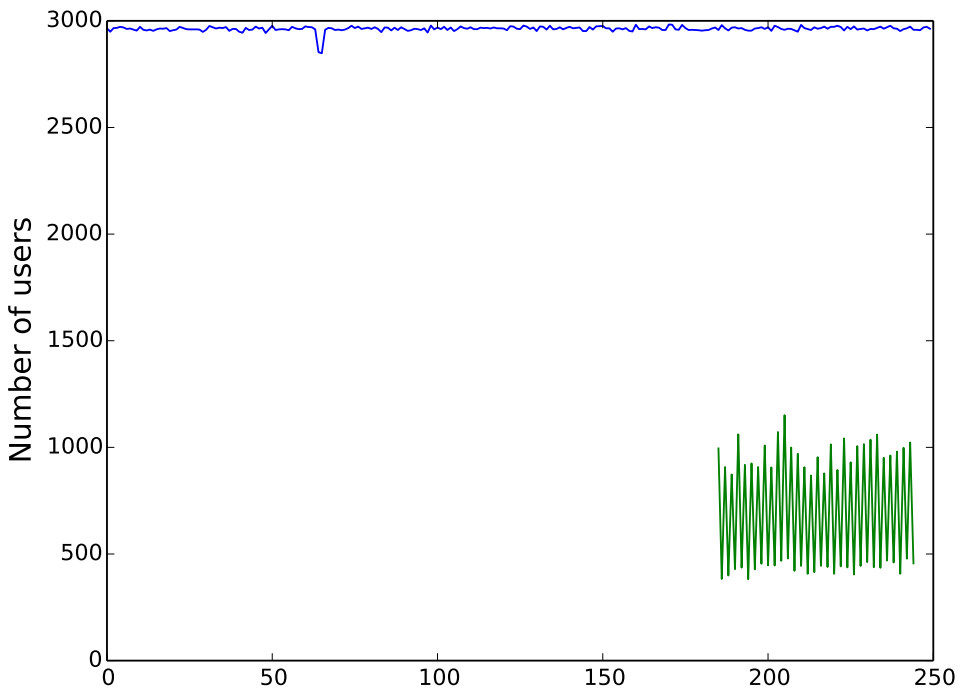

Figure B.47: Baseline - Uniques forecast from 2013-12-26 12:00:00 to 2014-01-25 00:00:00

σ (Real Data)	RMSE	MASE
7.22	2276.79	78.0702

Table B.47: Baseline - Error for Uniques forecast from 2013-12-26 12:00:00 to 2014-01-25 00:00:00

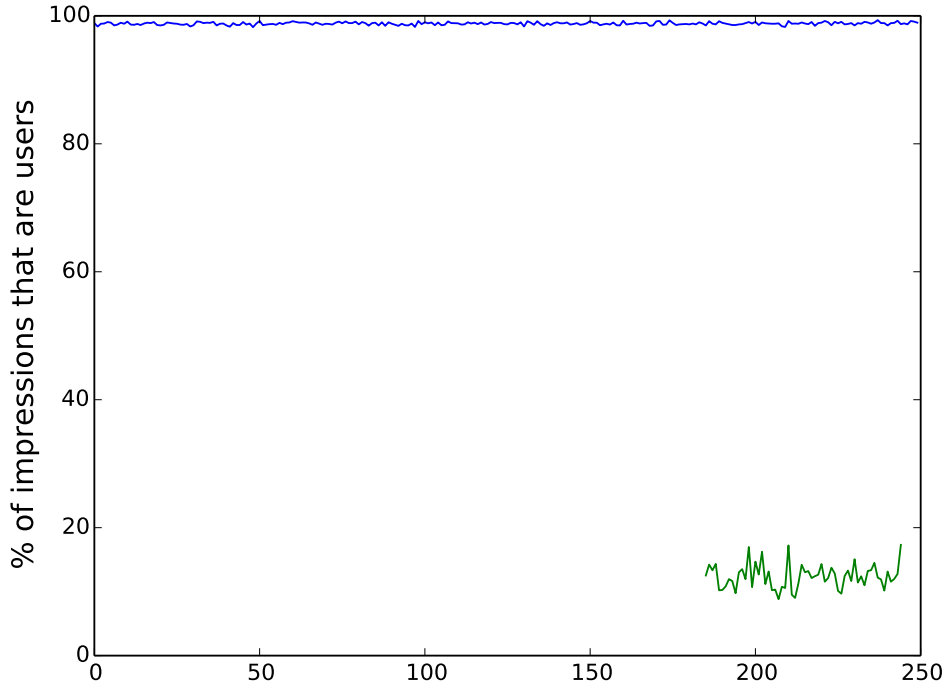

Figure B.48: Baseline - Uniques Percentage forecast from 2013-12-26 12:00:00 to 2014-01-25 00:00:00

σ (Real Data)	RMSE	MASE
0.22	86.47	113.3291

Table B.48: Baseline - Error for Uniques Percentage forecast from 2013-12-26 12:00:00 to 2014-01-25 00:00:00

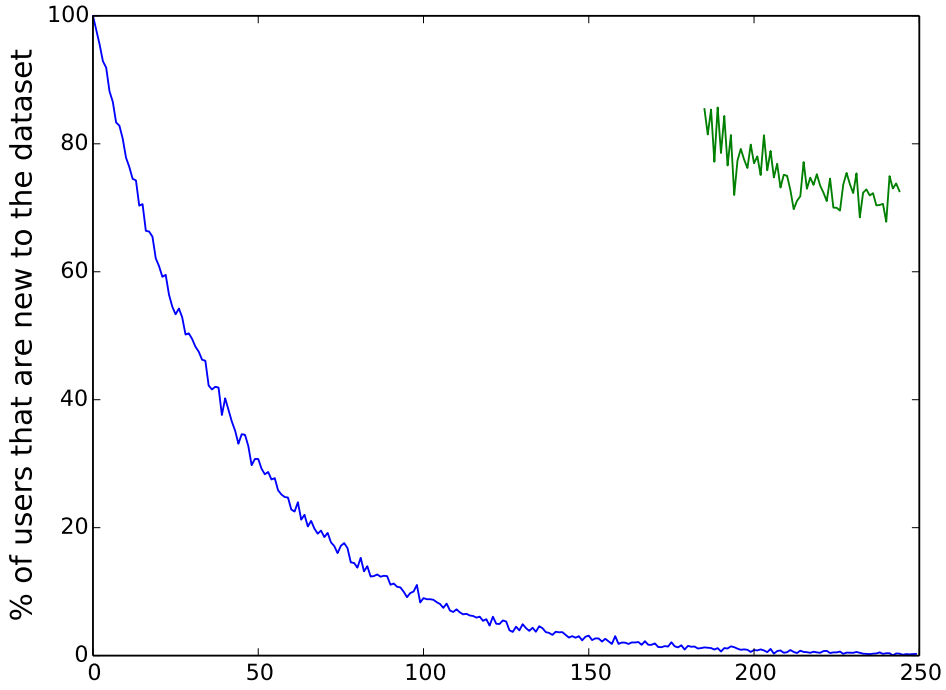

Figure B.49: Baseline - New uniques forecast from 2013-12-26 12:00:00 to 2014-01-25 00:00:00

σ (Real Data)	RMSE	MASE
0.33	74.47	26.7004

Table B.49: Baseline - Error for New Uniques forecast from 2013-12-26 12:00:00 to 2014-01-25 00:00:00

Case 2

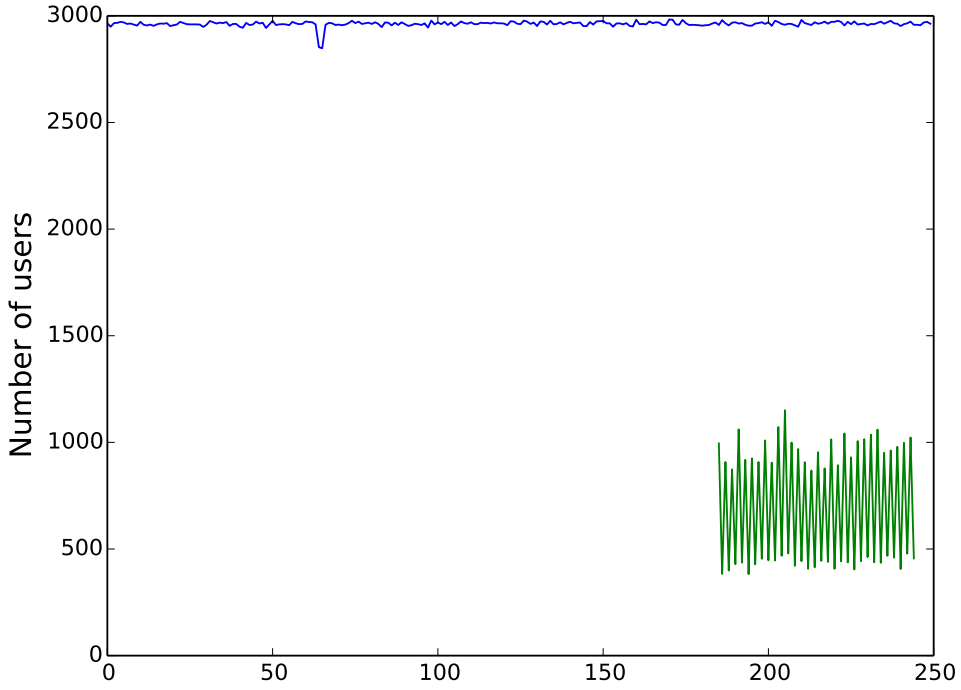

Figure B.50: Baseline - Uniques calculated using percentages forecast from 2013-12-26 12:00:00 to 2014-01-25 00:00:00

σ (Real Data)	RMSE	MASE
7.22	2276.96	78.076

Table B.50: Baseline - Error for Uniques calculated using percentages forecast from 2013-12-26 12:00:00 to 2014-01-25 00:00:00

B.11 Arima Allow Drift True - 12h

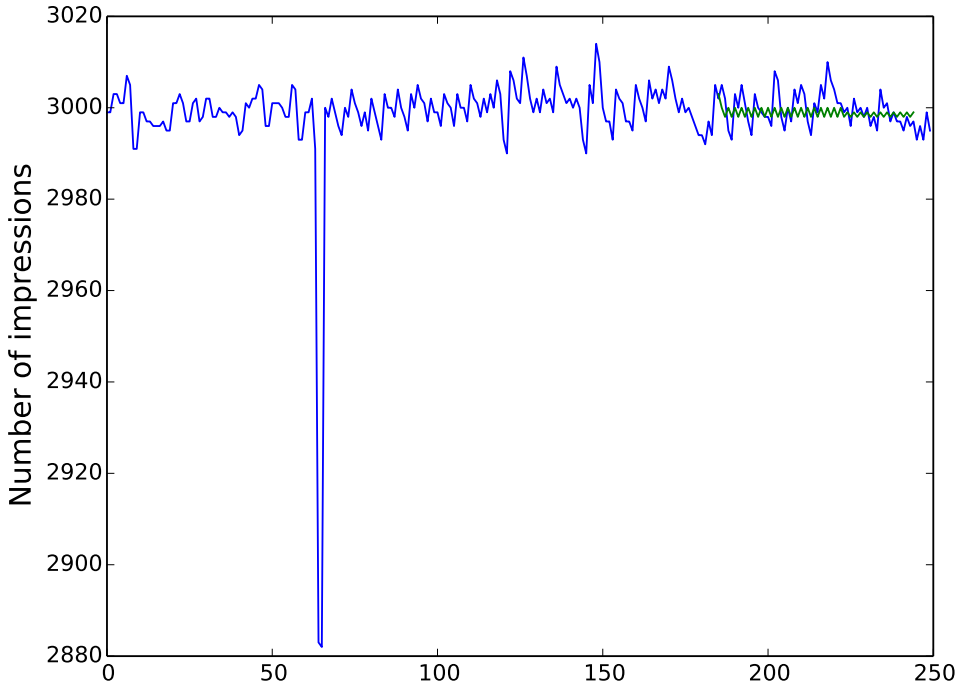

Figure B.51: Arima Allow Drift True - Impressions forecast from 2013-12-26 12:00:00 to 2014-01-25 00:00:00

σ (Real Data)	RMSE	MASE
3.79	3.54	0.1971

Table B.51: Arima Allow Drift True - Error for Impressions forecast from 2013-12-26 12:00:00 to 2014-01-25 00:00:00

Case 2

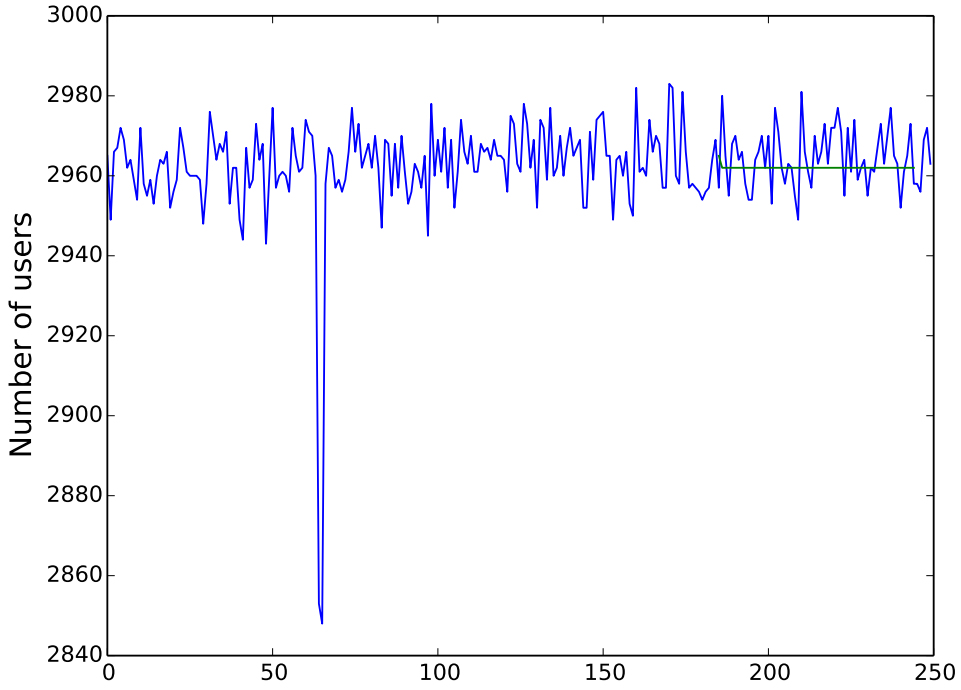

Figure B.52: Arima Allow Drift True - Uniques forecast from 2013-12-26 12:00:00 to 2014-01-25 00:00:00

σ (Real Data)	RMSE	MASE
7.22	7.79	0.213

Table B.52: Arima Allow Drift True - Error for Uniques forecast from 2013-12-26 12:00:00 to 2014-01-25 00:00:00

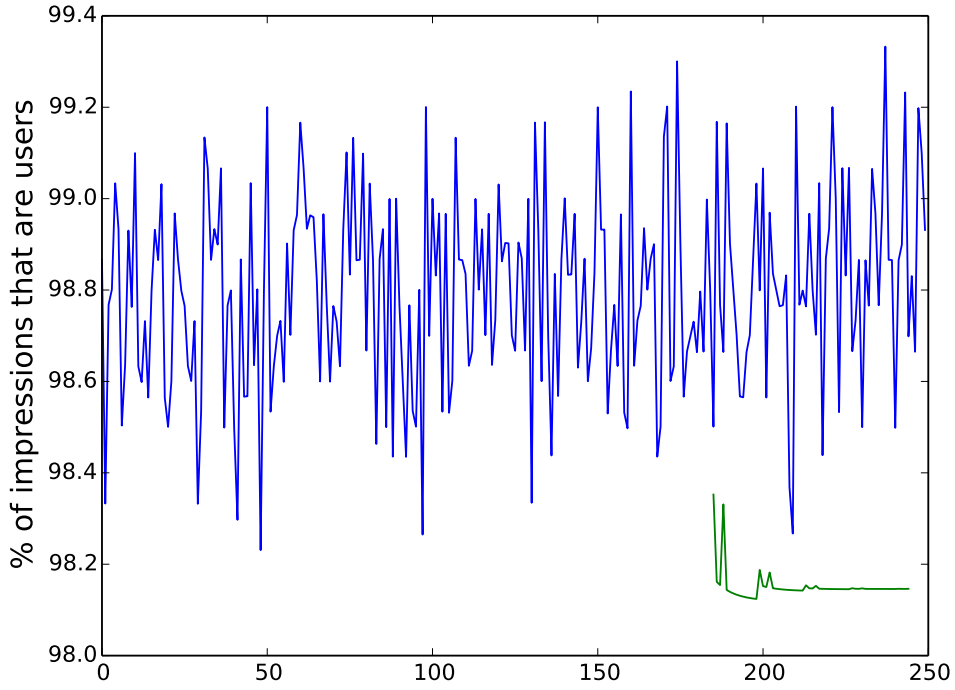

Figure B.53: Arima Allow Drift True - Uniques Percentage forecast from 2013-12-26 12:00:00 to 2014-01-25 00:00:00

σ (Real Data)	RMSE	MASE
0.22	0.71	0.8822

Table B.53: Arima Allow Drift True - Error for Uniques Percentage forecast from 2013-12-26 12:00:00 to 2014-01-25 00:00:00

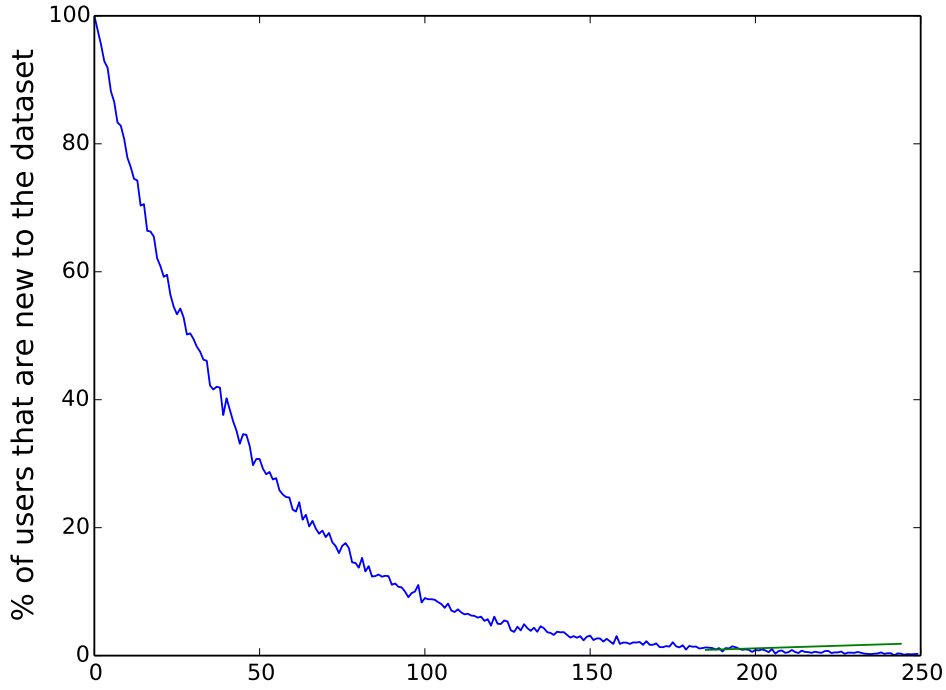

Figure B.54: Arima Allow Drift True - New uniques forecast from 2013-12-26 12:00:00 to 2014-01-25 00:00:00

σ (Real Data)	RMSE	MASE
0.33	0.92	0.2835

Table B.54: Arima Allow Drift True - Error for New Uniques forecast from 2013-12-26 12:00:00 to 2014-01-25 00:00:00

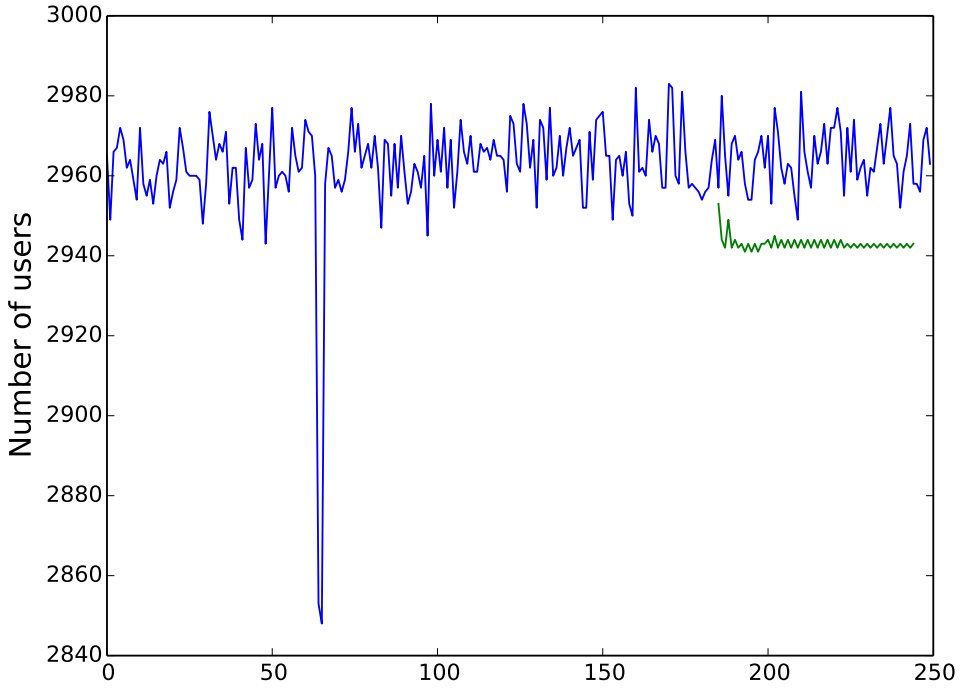

Figure B.55: Arima Allow Drift True - Uniques calculated using percentages forecast from 2013-12-26 12:00:00 to 2014-01-25 00:00:00

σ (Real Data)	RMSE	MASE
7.22	22.9	0.7467

Table B.55: Arima Allow Drift True - Error for Uniques calculated using percentages forecast from 2013-12-26 12:00:00 to 2014-01-25 00:00:00

B.12 Arima Allow Drift False - 12h

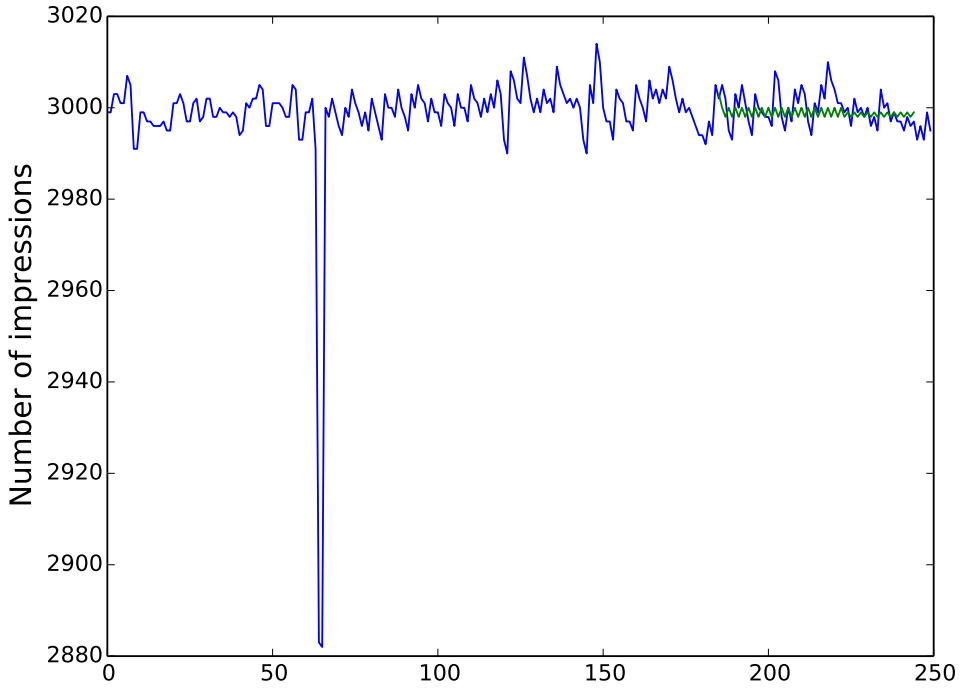

Figure B.56: Arima Allow Drift False - Impressions forecast from 2013-12-26 12:00:00 to 2014-01-25 00:00:00

σ (Real Data)	RMSE	MASE
3.79	3.54	0.1971

Table B.56: Arima Allow Drift False - Error for Impressions forecast from 2013-12-26 12:00:00 to 2014-01-25 00:00:00

Case 2

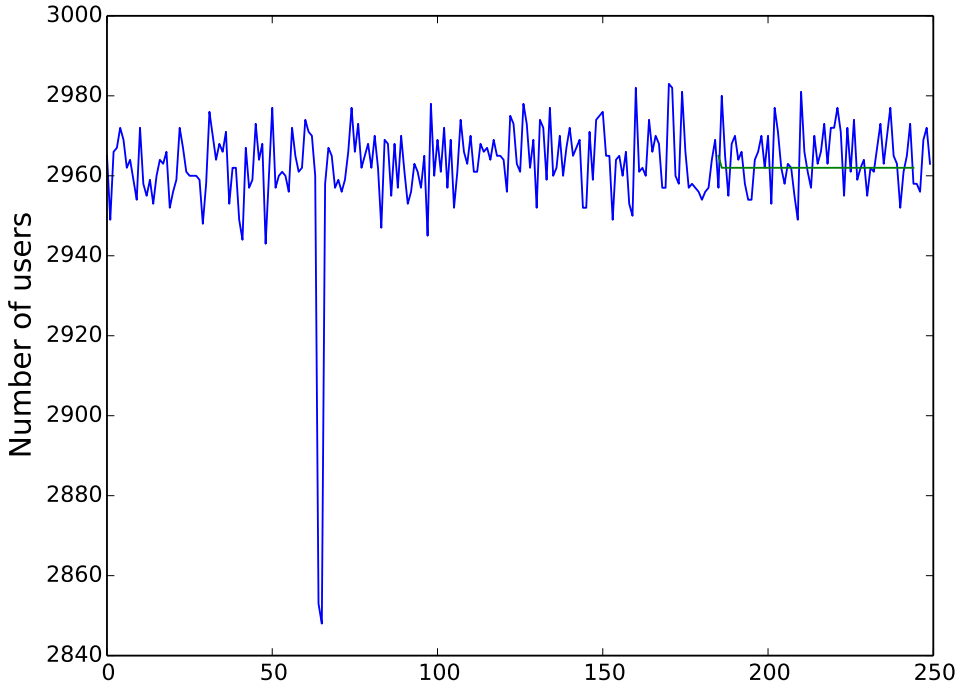

Figure B.57: Arima Allow Drift False - Uniques forecast from 2013-12-26 12:00:00 to 2014-01-25 00:00:00

σ (Real Data)	RMSE	MASE
7.22	7.79	0.213

Table B.57: Arima Allow Drift False - Error for Uniques forecast from 2013-12-26 12:00:00 to 2014-01-25 00:00:00

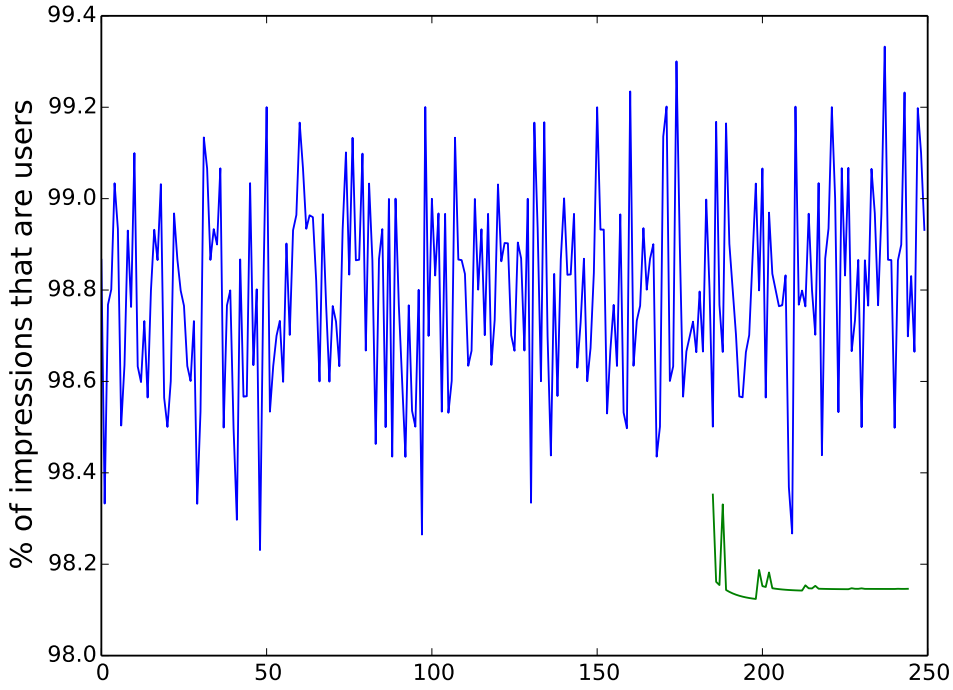

Figure B.58: Arima Allow Drift False - Uniques Percentage forecast from 2013-12-26 12:00:00 to 2014-01-25 00:00:00

σ (Real Data)	RMSE	MASE
0.22	0.71	0.8822

Table B.58: Arima Allow Drift False - Error for Uniques Percentage forecast from 2013-12-26 12:00:00 to 2014-01-25 00:00:00

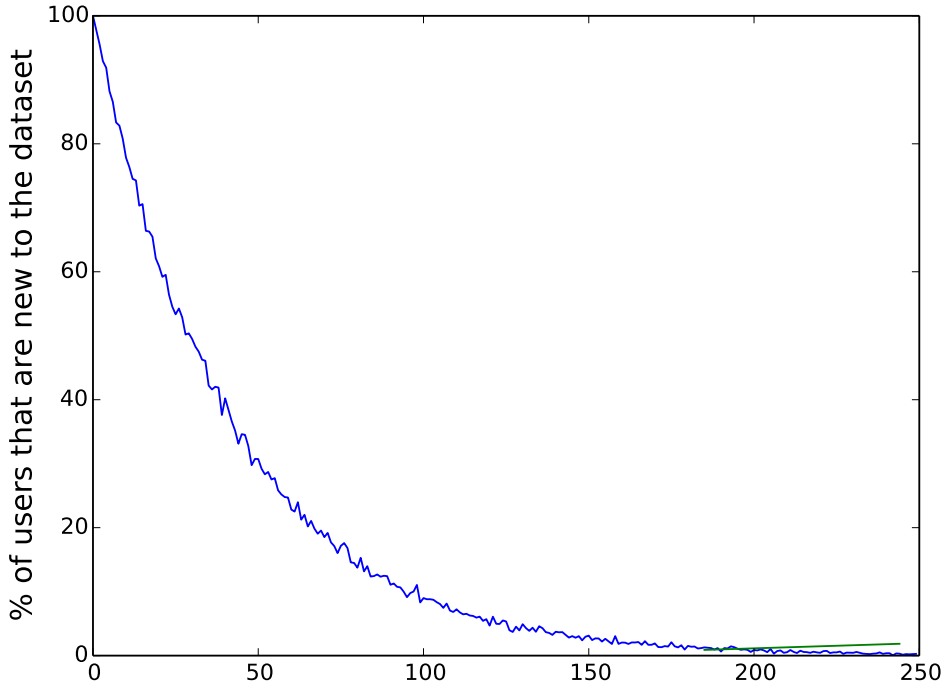

Figure B.59: Arima Allow Drift False - New uniques forecast from 2013-12-26 12:00:00 to 2014-01-25 00:00:00

σ (Real Data)	RMSE	MASE
0.33	0.92	0.2835

Table B.59: Arima Allow Drift False - Error for New Uniques forecast from 2013-12-26 12:00:00 to 2014-01-25 00:00:00

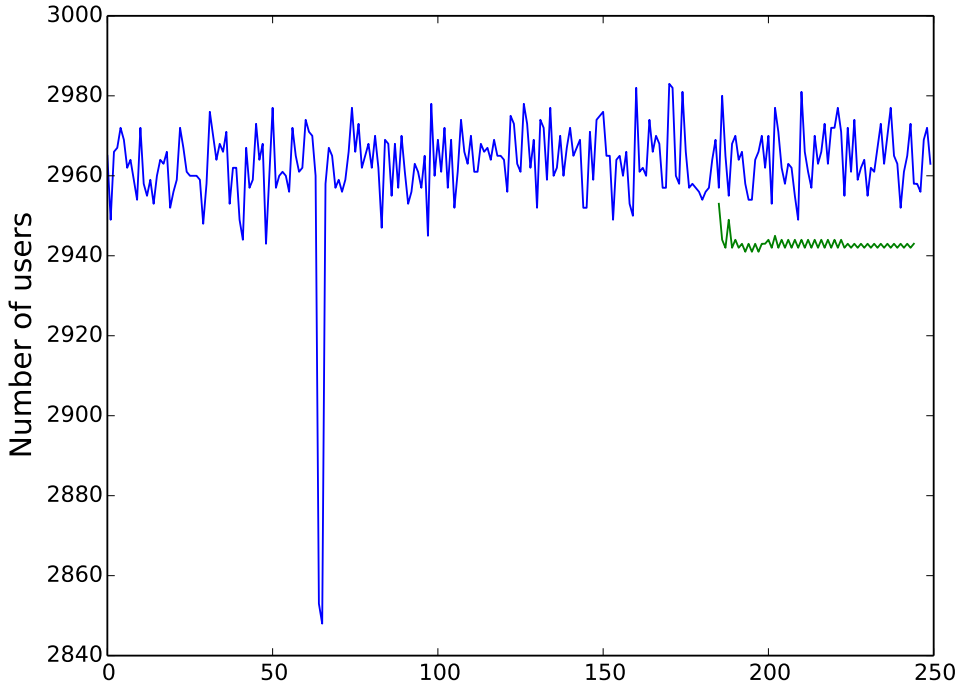

Figure B.60: Arima Allow Drift False - Uniques calculated using percentages forecast from 2013-12-26 12:00:00 to 2014-01-25 00:00:00

σ (Real Data)	RMSE	MASE
7.22	22.9	0.7467

Table B.60: Arima Allow Drift False - Error for Uniques calculated using percentages forecast from 2013-12-26 12:00:00 to 2014-01-25 00:00:00

B.13 Baseline - 24h

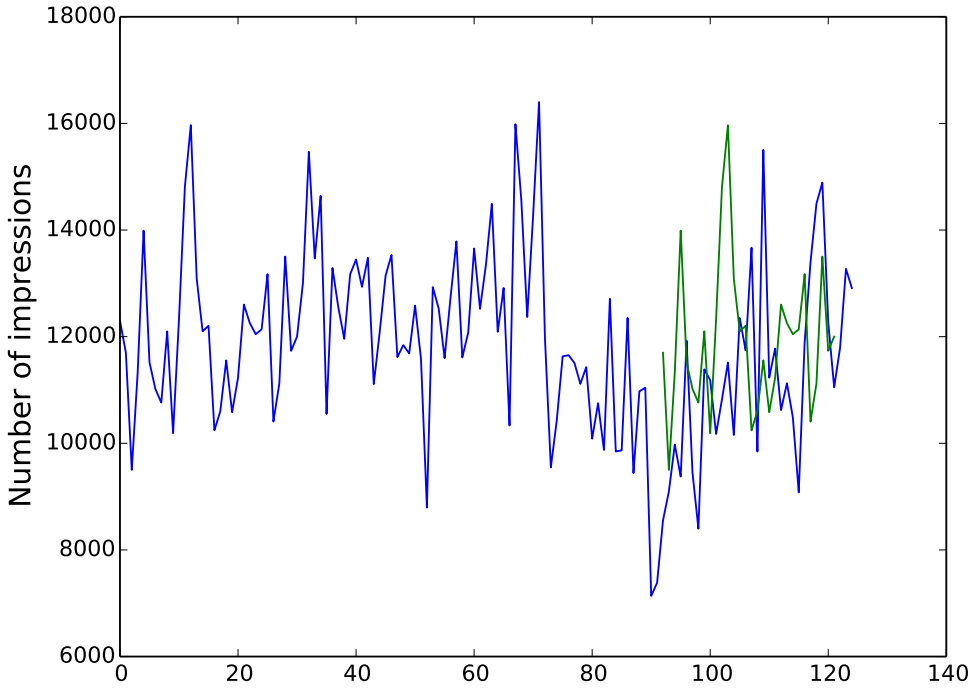

Figure B.61: Baseline - Impressions forecast from 2013-12-26 00:00:00 to 2014-01-24 00:00:00

σ (Real Data)	RMSE	MASE
1752.65	2329.76	0.4198

Table B.61: Baseline - Error for Impressions forecast from 2013-12-26 00:00:00 to 2014-01-24 00:00:00

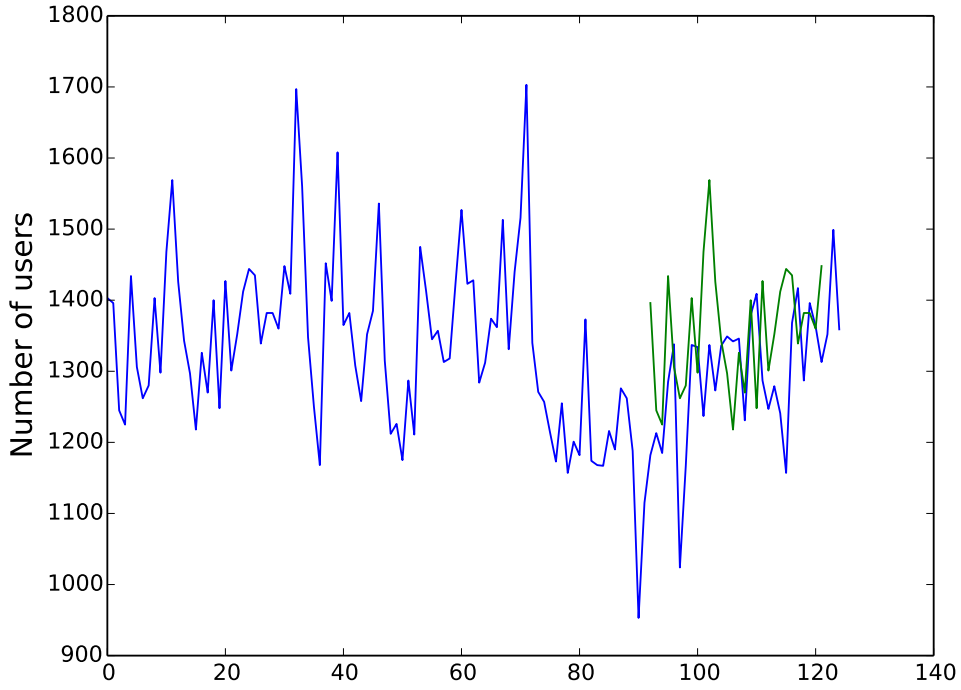

Figure B.62: Baseline - Uniques forecast from 2013-12-26 00:00:00 to 2014-01-24 00:00:00

σ (Real Data)	RMSE	MASE
91.16	129.25	0.3395

Table B.62: Baseline - Error for Uniques forecast from 2013-12-26 00:00:00 to 2014-01-24 00:00:00

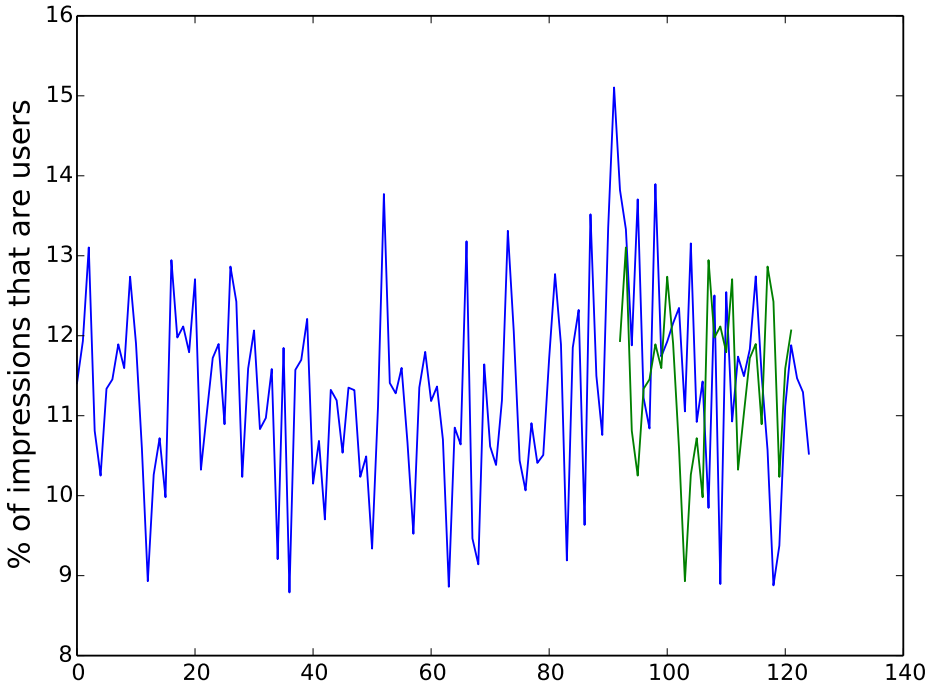

Figure B.63: Baseline - Uniques Percentage forecast from 2013-12-26 00:00:00 to 2014-01-24 00:00:00

σ (Real Data)	RMSE	MASE
1.25	1.69	0.3407

Table B.63: Baseline - Error for Uniques Percentage forecast from 2013-12-26 00:00:00 to 2014-01-24 00:00:00

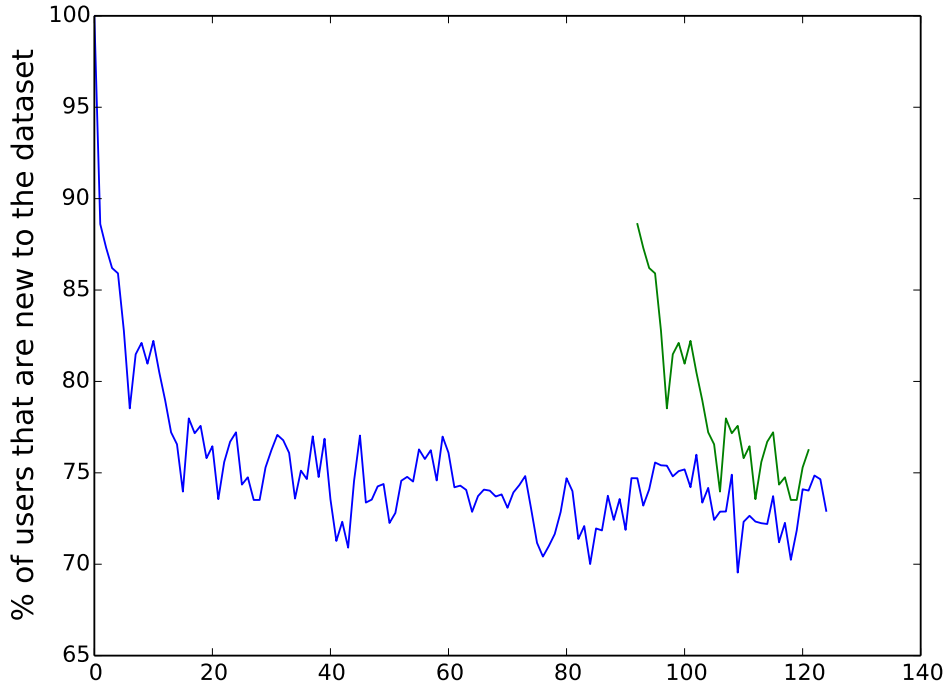

Figure B.64: Baseline - New uniques forecast from 2013-12-26 00:00:00 to 2014-01-24 00:00:00

σ (Real Data)	RMSE	MASE
1.54	6.28	1.118

Table B.64: Baseline - Error for New Uniques forecast from 2013-12-26 00:00:00 to 2014-01-24 00:00:00

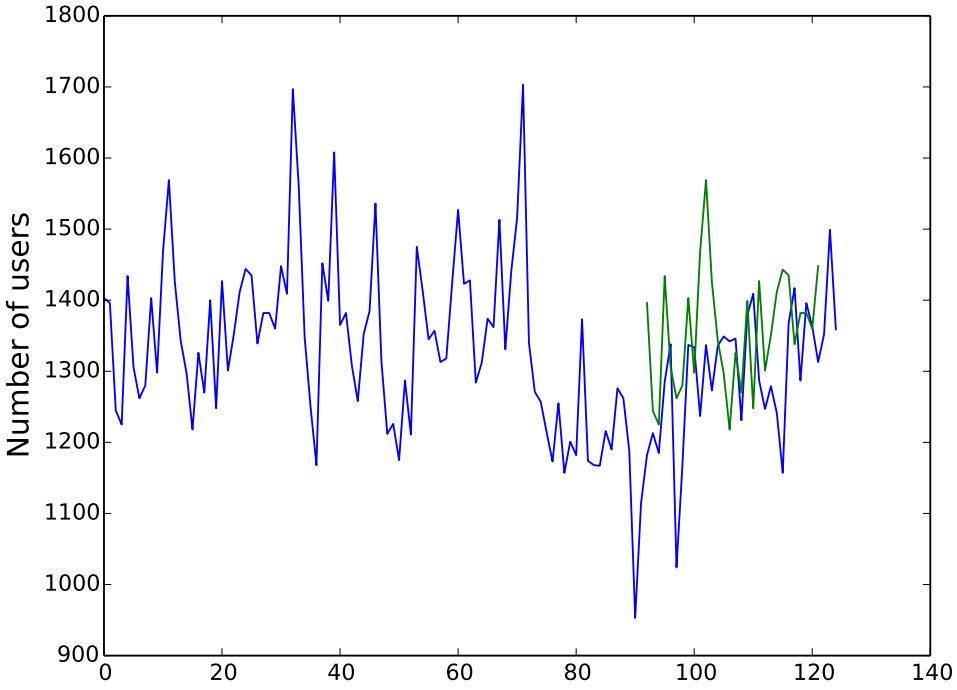

Figure B.65: Baseline - Uniques calculated using percentages forecast from 2013-12-26 00:00:00 to 2014-01-24 00:00:00

σ (Real Data)	RMSE	MASE
91.16	129.13	0.3391

Table B.65: Baseline - Error for Uniques calculated using percentages forecast from 2013-12-26 00:00:00 to 2014-01-24 00:00:00

B.14 Arima Allow Drift True - 24h

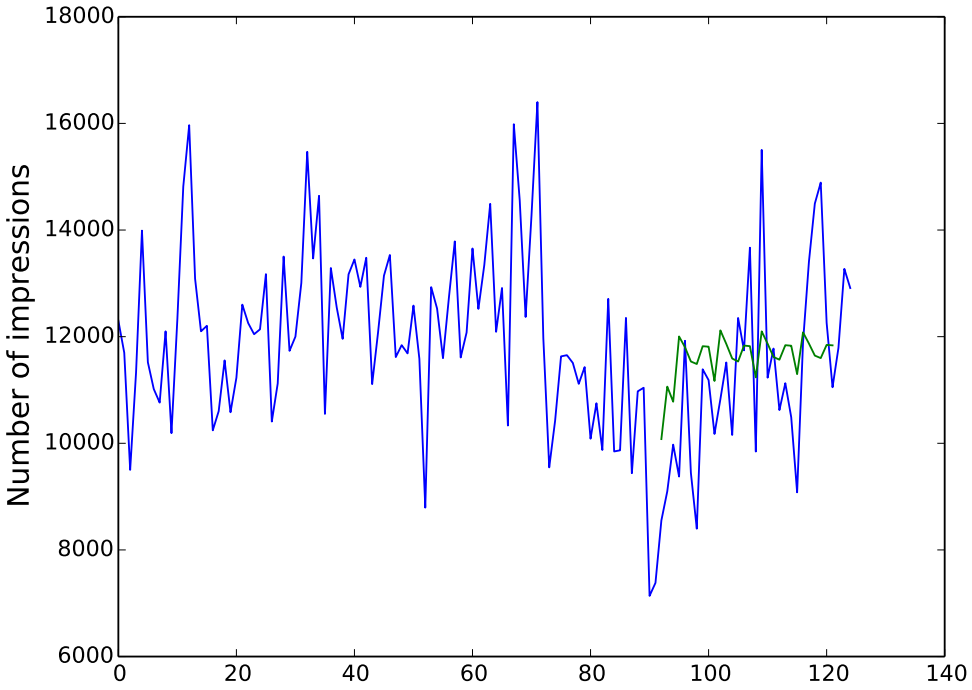

Figure B.66: Arima Allow Drift True - Impressions forecast from 2013-12-26 00:00:00 to 2014-01-24 00:00:00

σ (Real Data)	RMSE	MASE
1752.65	1646.92	0.2925

Table B.66: Arima Allow Drift True - Error for Impressions forecast from 2013-12-26 00:00:00 to 2014-01-24 00:00:00

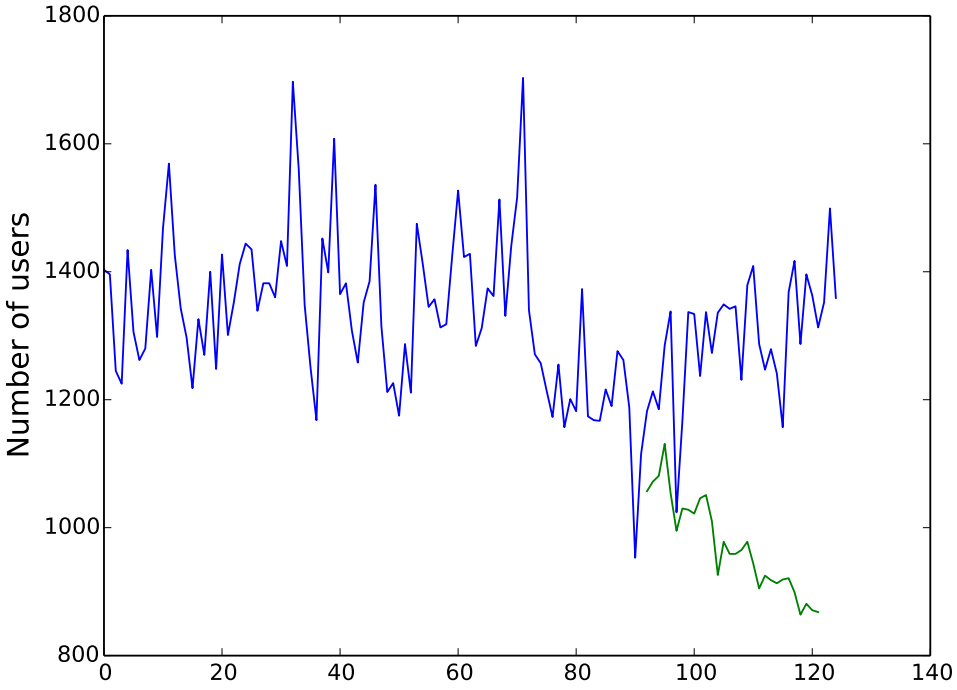

Figure B.67: Arima Allow Drift True - Uniques forecast from 2013-12-26 00:00:00 to 2014-01-24 00:00:00

σ (Real Data)	RMSE	MASE
91.16	341.42	1.0471

Table B.67: Arima Allow Drift True - Error for Uniques forecast from 2013-12-26 00:00:00 to 2014-01-24 00:00:00

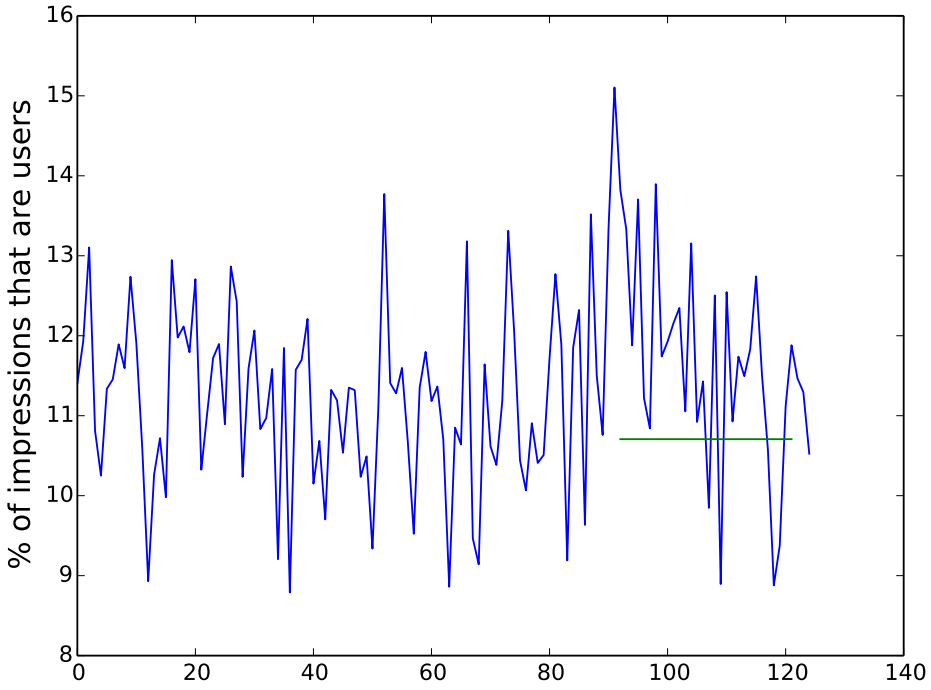

Figure B.68: Arima Allow Drift True - Uniques Percentage forecast from 2013-12-26 00:00:00 to 2014-01-24 00:00:00

σ (Real Data)	RMSE	MASE
1.25	1.6	0.348

Table B.68: Arima Allow Drift True - Error for Uniques Percentage forecast from 2013-12-26 00:00:00 to 2014-01-24 00:00:00

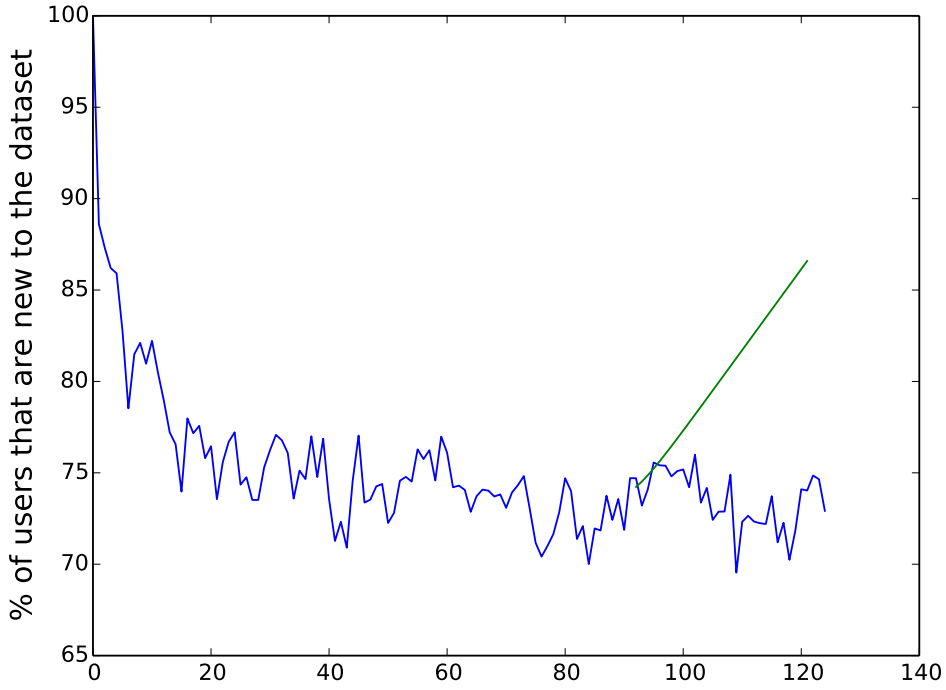

Figure B.69: Arima Allow Drift True - New uniques forecast from 2013-12-26 00:00:00 to 2014-01-24 00:00:00

σ (Real Data)	RMSE	MASE
1.54	8.36	1.47

Table B.69: Arima Allow Drift True - Error for New Uniques forecast from 2013-12-26 00:00:00 to 2014-01-24 00:00:00

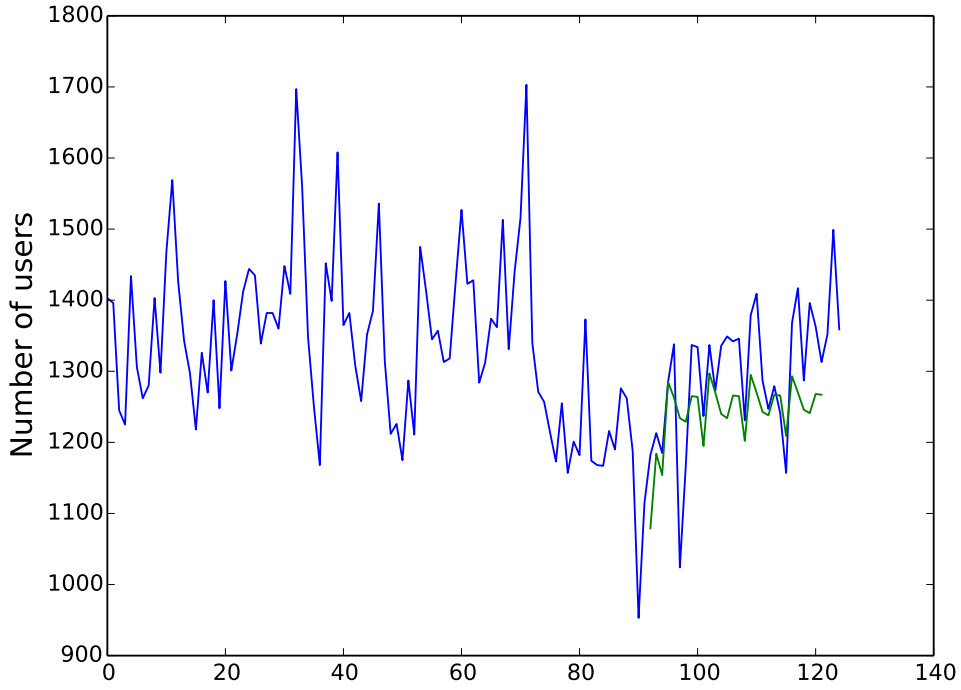

Figure B.70: Arima Allow Drift True - Uniques calculated using percentages forecast from 2013-12-26 00:00:00 to 2014-01-24 00:00:00

σ (Real Data)	RMSE	MASE
91.16	83.87	0.2271

Table B.70: Arima Allow Drift True - Error for Uniques calculated using percentages forecast from 2013-12-26 00:00:00 to 2014-01-24 00:00:00

B.15 Arima Allow Drift False - 24h

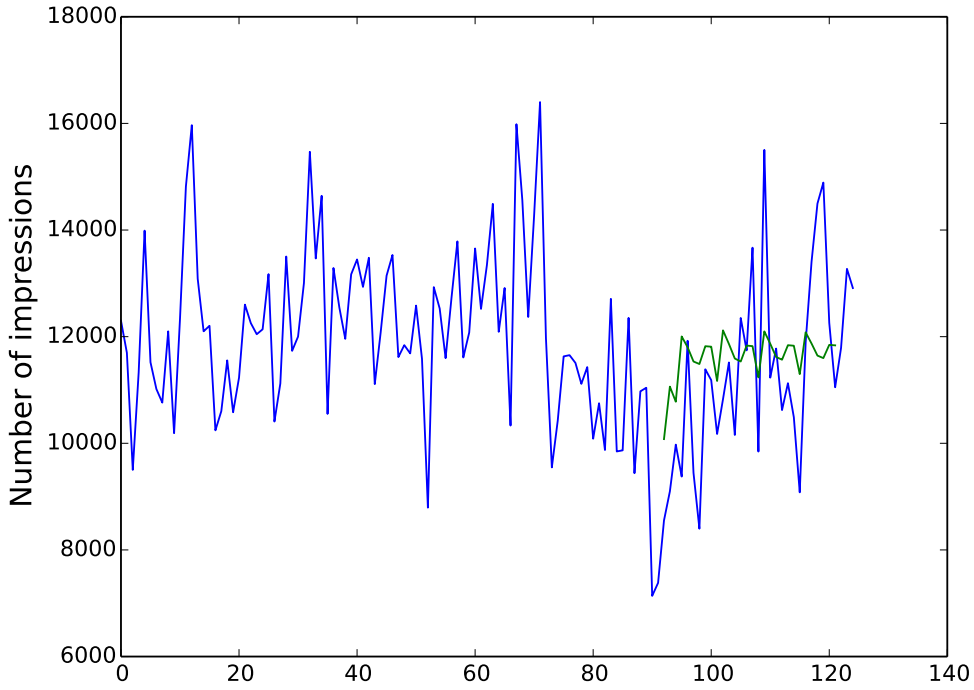

Figure B.71: Arima Allow Drift False - Impressions forecast from 2013-12-26 00:00:00 to 2014-01-24 00:00:00

σ (Real Data)	RMSE	MASE
1752.65	1646.92	0.2925

Table B.71: Arima Allow Drift False - Error for Impressions forecast from 2013-12-26 00:00:00 to 2014-01-24 00:00:00

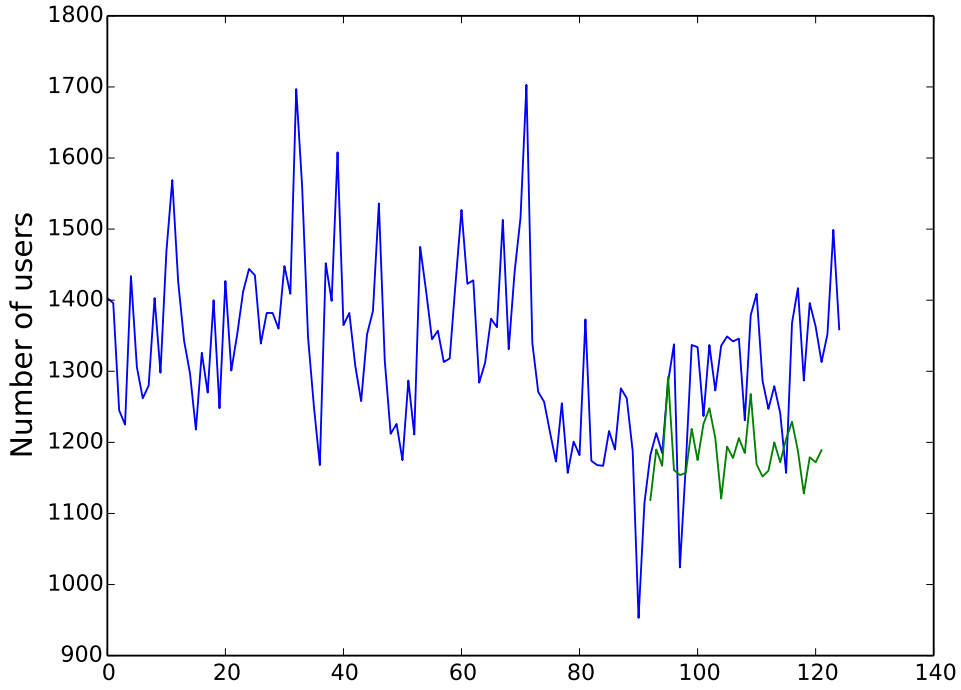

Figure B.72: Arima Allow Drift False - Uniques forecast from 2013-12-26 00:00:00 to 2014-01-24 00:00:00

σ (Real Data)	RMSE	MASE
91.16	132.77	0.3774

Table B.72: Arima Allow Drift False - Error for Uniques forecast from 2013-12-26 00:00:00 to 2014-01-24 00:00:00

Case 2

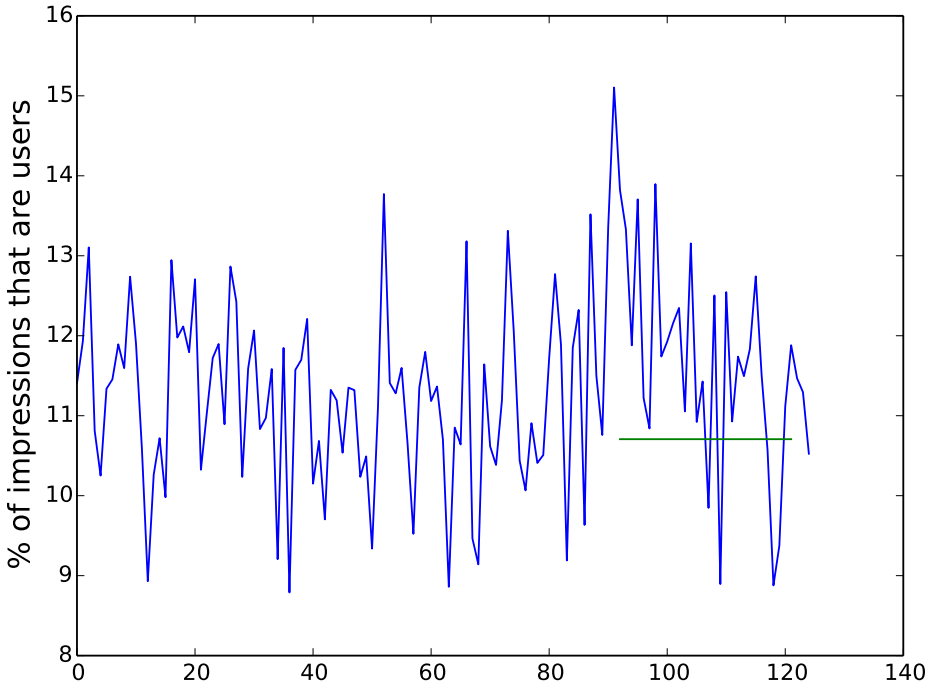

Figure B.73: Arima Allow Drift False - Uniques Percentage forecast from 2013-12-26 00:00:00 to 2014-01-24 00:00:00

σ (Real Data)	RMSE	MASE
1.25	1.6	0.348

Table B.73: Arima Allow Drift False - Error for Uniques Percentage forecast from 2013-12-26 00:00:00 to 2014-01-24 00:00:00

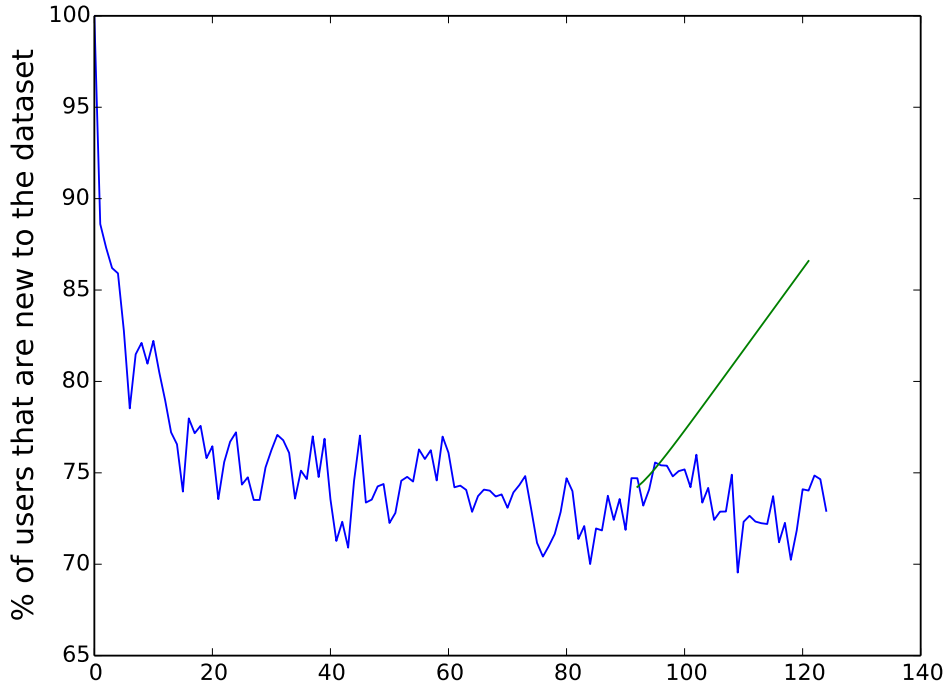

Figure B.74: Arima Allow Drift False - New uniques forecast from 2013-12-26 00:00:00 to 2014-01-24 00:00:00

σ (Real Data)	RMSE	MASE
1.54	8.36	1.47

Table B.74: Arima Allow Drift False - Error for New Uniques forecast from 2013-12-26 00:00:00 to 2014-01-24 00:00:00

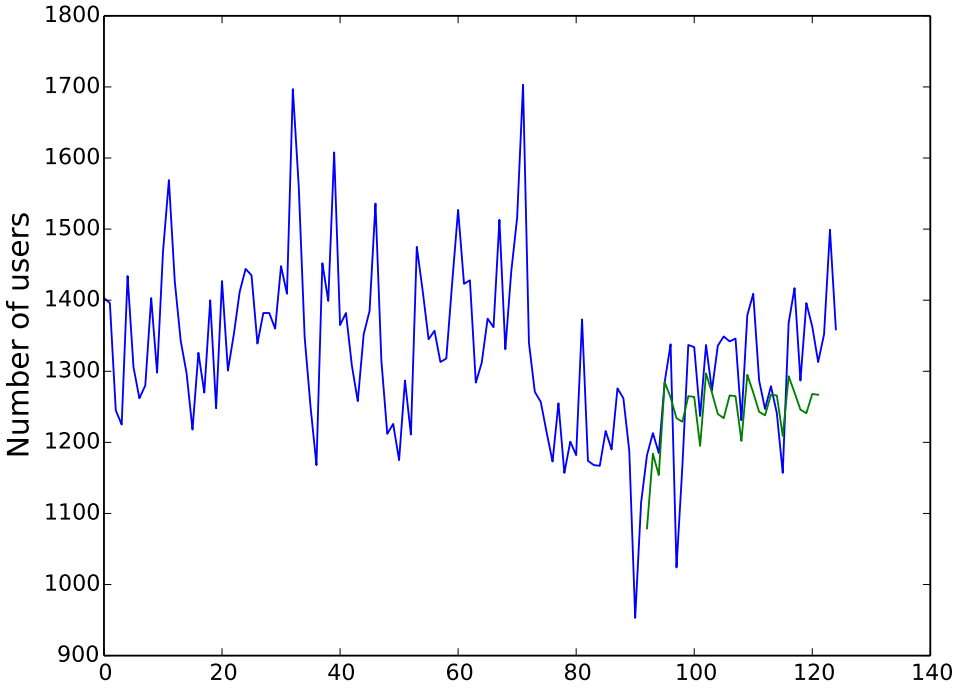

Figure B.75: Arima Allow Drift False - Uniques calculated using percentages forecast from 2013-12-26 00:00:00 to 2014-01-24 00:00:00

σ (Real Data)	RMSE	MASE
91.16	83.87	0.2271

Table B.75: Arima Allow Drift False - Error for Uniques calculated using percentages forecast from 2013-12-26 00:00:00 to 2014-01-24 00:00:00

Appendix C

Case 3

C.1 Baseline - 4h

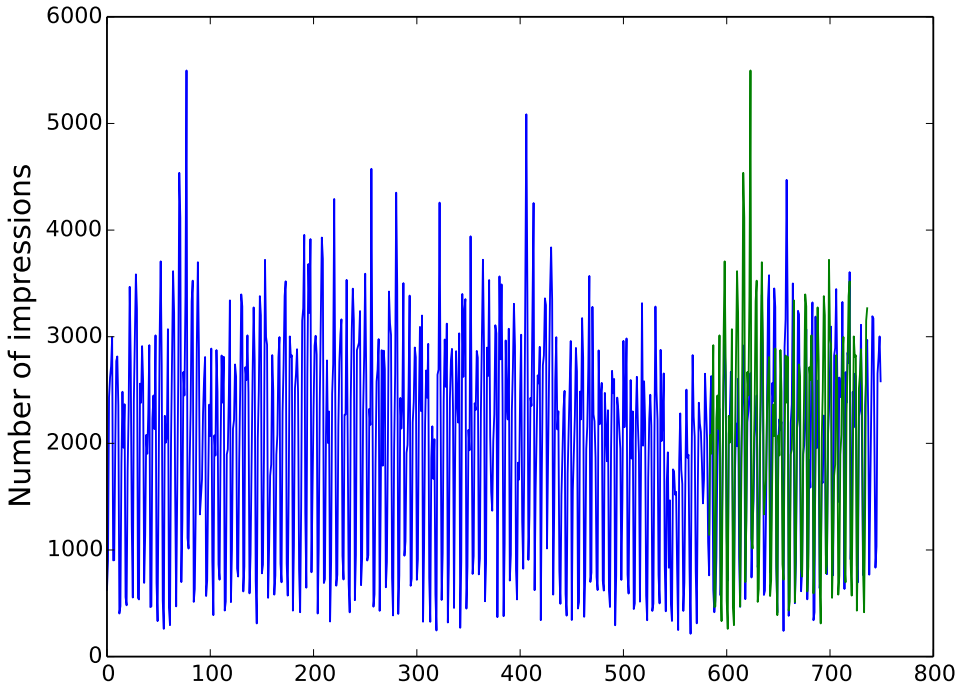

Figure C.1: Baseline - Impressions forecast from 2013-12-31 04:00:00 to 2014-01-25 16:00:00

Case 3

σ (Real Data)	RMSE	MASE
929.73	737.22	0.1548

Table C.1: Baseline - Error for Impressions forecast from 2013-12-31 04:00:00 to 2014-01-25 16:00:00

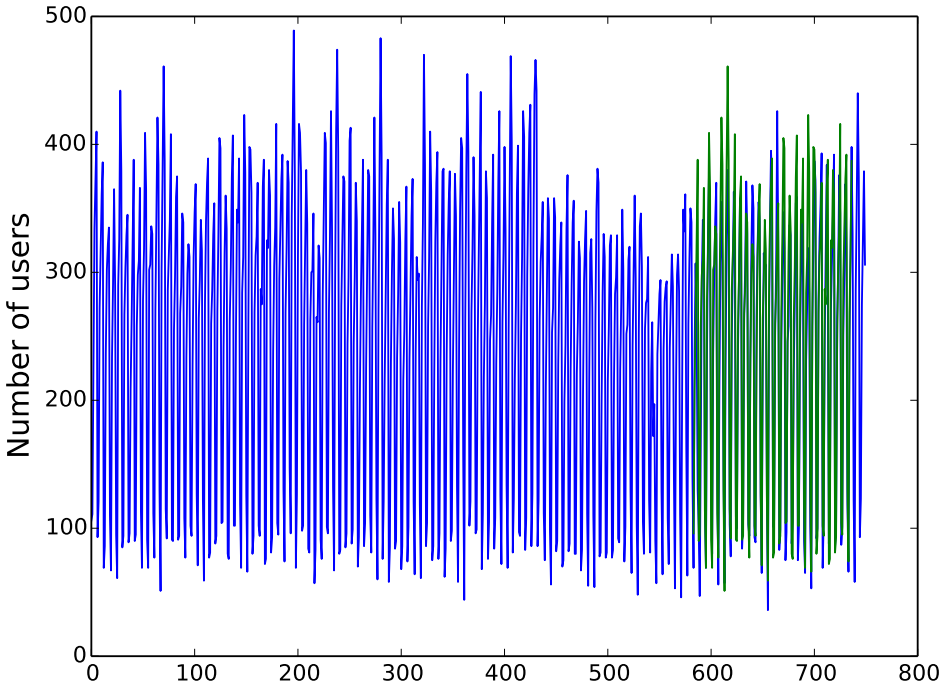

Figure C.2: Baseline - Uniques forecast from 2013-12-31 04:00:00 to 2014-01-25 16:00:00

σ (Real Data)	RMSE	MASE
112.4	42.96	0.0774

Table C.2: Baseline - Error for Uniques forecast from 2013-12-31 04:00:00 to 2014-01-25 16:00:00

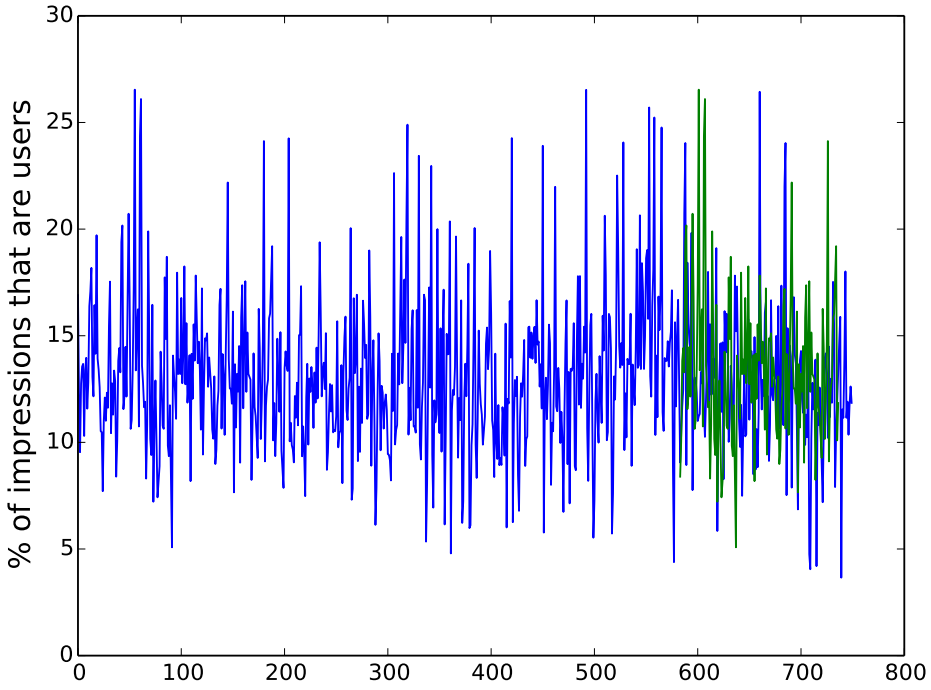

Figure C.3: Baseline - Uniques Percentage forecast from 2013-12-31 04:00:00 to 2014-01-25 16:00:00

σ (Real Data)	RMSE	MASE
3.4	5.07	0.2827

Table C.3: Baseline - Error for Uniques Percentage forecast from 2013-12-31 04:00:00 to 2014-01-25 16:00:00

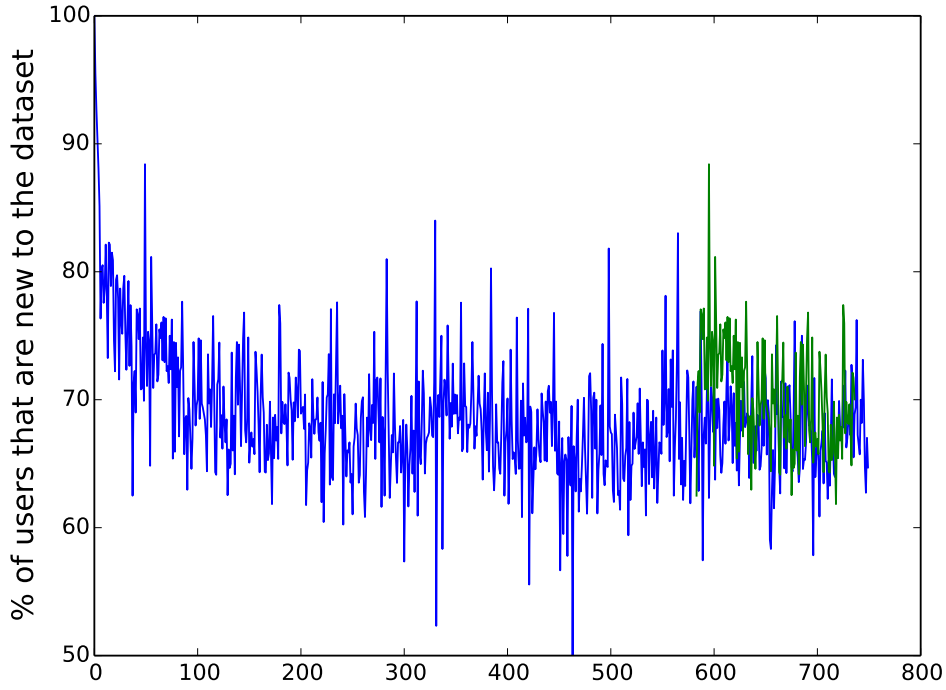

Figure C.4: Baseline - New uniques forecast from 2013-12-31 04:00:00 to 2014-01-25 16:00:00

σ (Real Data)	RMSE	MASE
3.37	5.68	0.2762

Table C.4: Baseline - Error for New Uniques forecast from 2013-12-31 04:00:00 to 2014-01-25 16:00:00

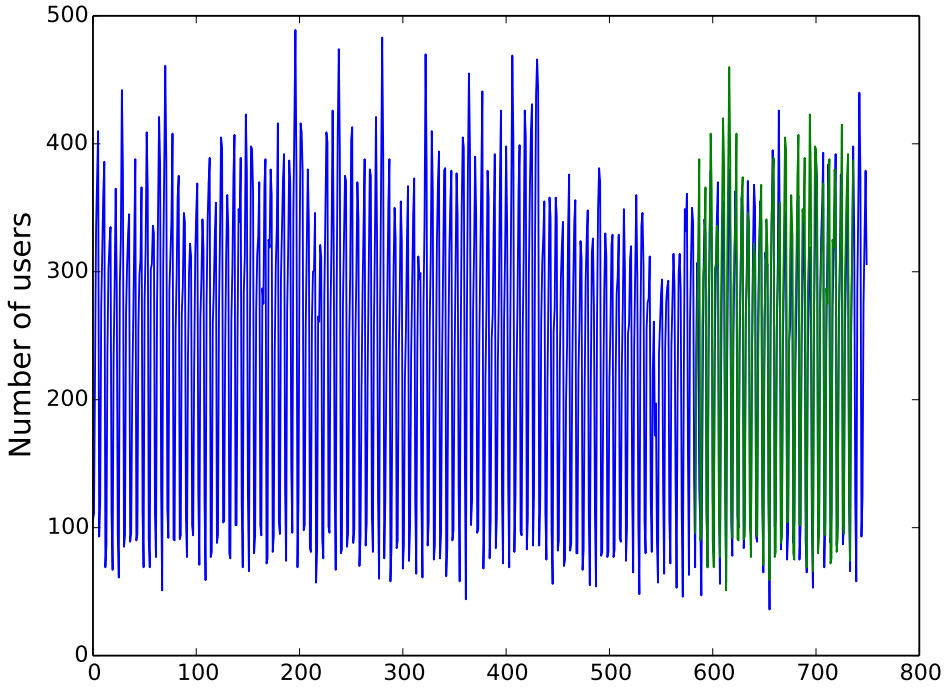

Figure C.5: Baseline - Uniques calculated using percentages forecast from 2013-12-31 04:00:00 to 2014-01-25 16:00:00

σ (Real Data)	RMSE	MASE
112.4	42.91	0.0773

Table C.5: Baseline - Error for Uniques calculated using percentages forecast from 2013-12-31 04:00:00 to 2014-01-25 16:00:00

C.2 Arima Allow Drift True - 4h

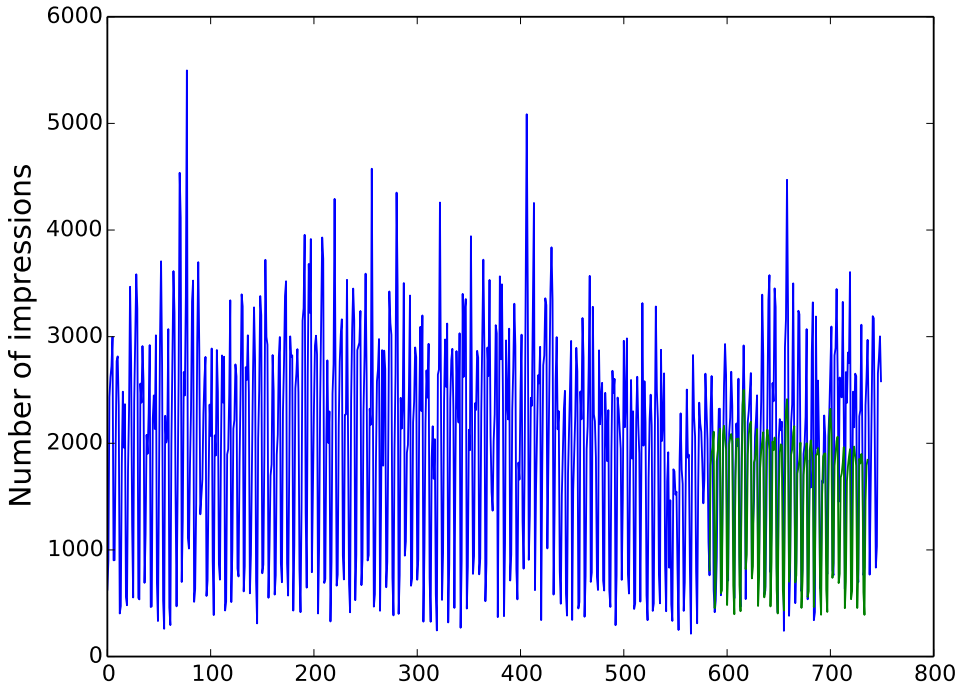

Figure C.6: Arima Allow Drift True - Impressions forecast from 2013-12-31 04:00:00 to 2014-01-25 16:00:00

σ (Real Data)	RMSE	MASE
929.73	692.61	0.1441

Table C.6: Arima Allow Drift True - Error for Impressions forecast from 2013-12-31 04:00:00 to 2014-01-25 16:00:00

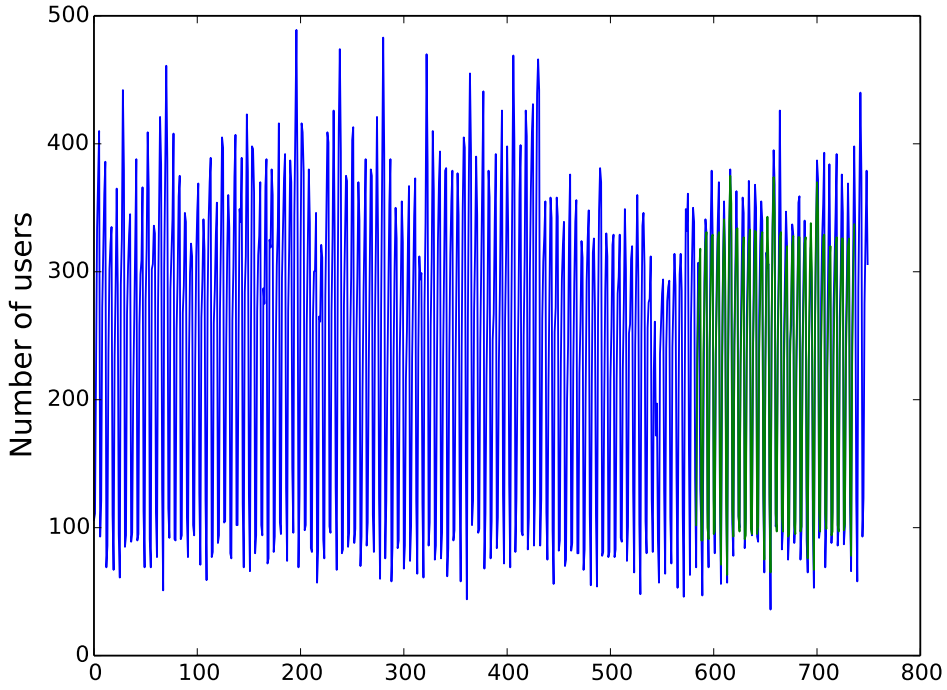

Figure C.7: Arima Allow Drift True - Uniques forecast from 2013-12-31 04:00:00 to 2014-01-25 16:00:00

σ (Real Data)	RMSE	MASE
112.4	30.3	0.0558

Table C.7: Arima Allow Drift True - Error for Uniques forecast from 2013-12-31 04:00:00 to 2014-01-25 16:00:00

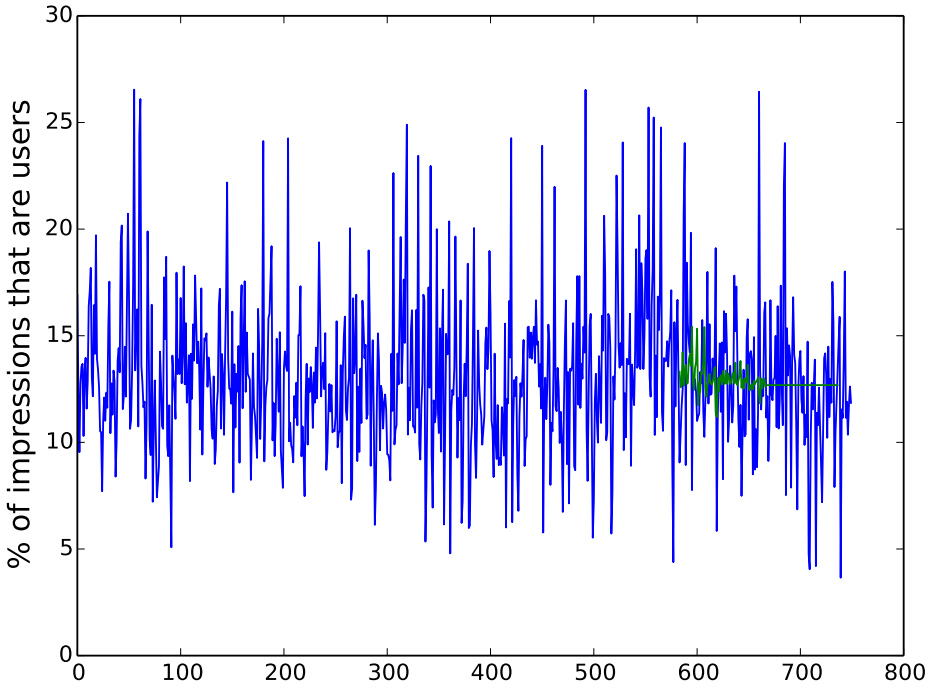

Figure C.8: Arima Allow Drift True - Uniques Percentage forecast from 2013-12-31 04:00:00 to 2014-01-25 16:00:00

σ (Real Data)	RMSE	MASE
3.4	3.49	0.1947

Table C.8: Arima Allow Drift True - Error for Uniques Percentage forecast from 2013-12-31 04:00:00 to 2014-01-25 16:00:00

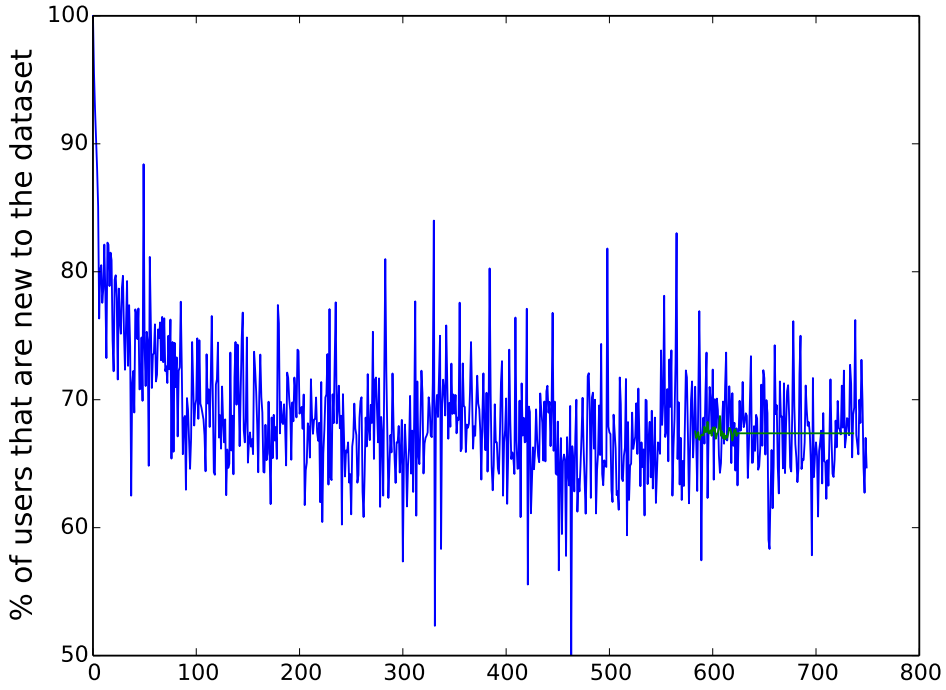

Figure C.9: Arima Allow Drift True - New uniques forecast from 2013-12-31 04:00:00 to 2014-01-25 16:00:00

σ (Real Data)	RMSE	MASE
3.37	3.38	0.1695

Table C.9: Arima Allow Drift True - Error for New Uniques forecast from 2013-12-31 04:00:00 to 2014-01-25 16:00:00

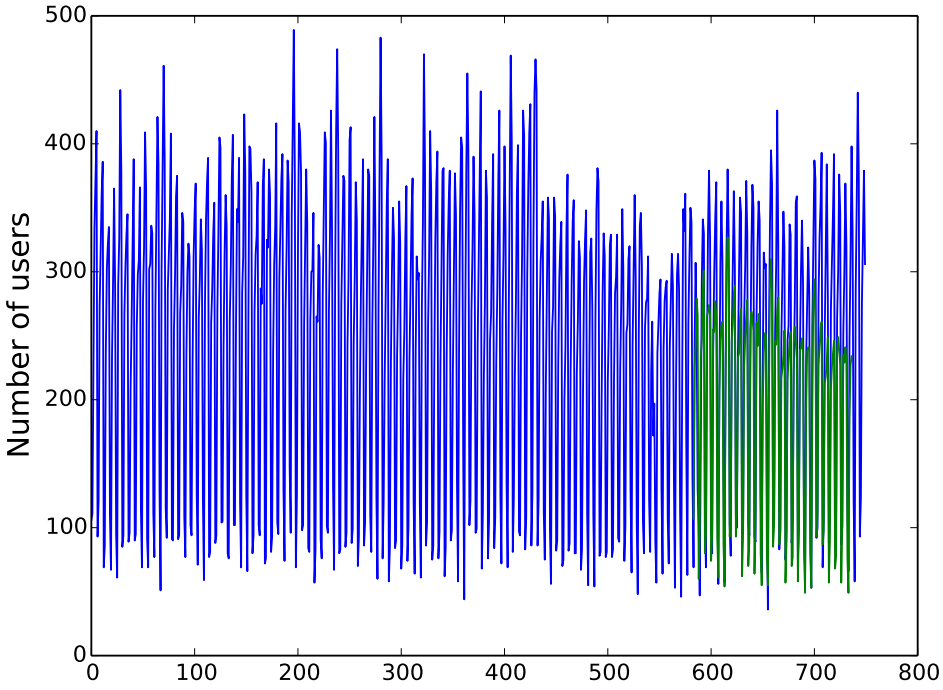

Figure C.10: Arima Allow Drift True - Uniques calculated using percentages forecast from 2013-12-31 04:00:00 to 2014-01-25 16:00:00

σ (Real Data)	RMSE	MASE
112.4	63.21	0.1307

Table C.10: Arima Allow Drift True - Error for Uniques calculated using percentages forecast from 2013-12-31 04:00:00 to 2014-01-25 16:00:00

C.3 Arima Allow Drift False - 4h

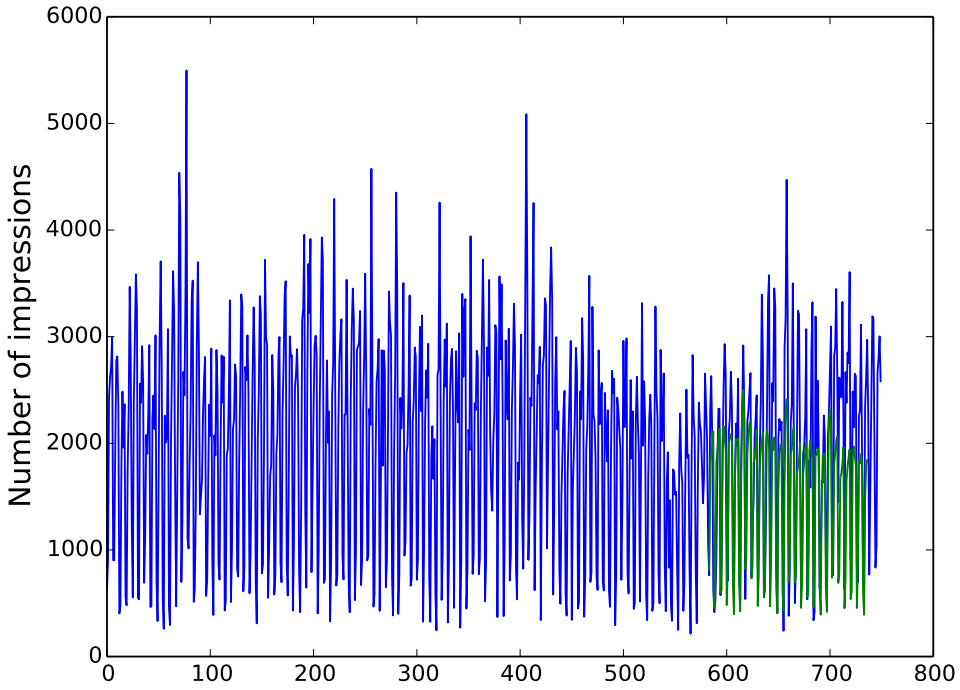

Figure C.11: Arima Allow Drift False - Impressions forecast from 2013-12-31 04:00:00 to 2014-01-25 16:00:00

σ (Real Data)	RMSE	MASE
929.73	692.61	0.1441

Table C.11: Arima Allow Drift False - Error for Impressions forecast from 2013-12-31 04:00:00 to 2014-01-25 16:00:00

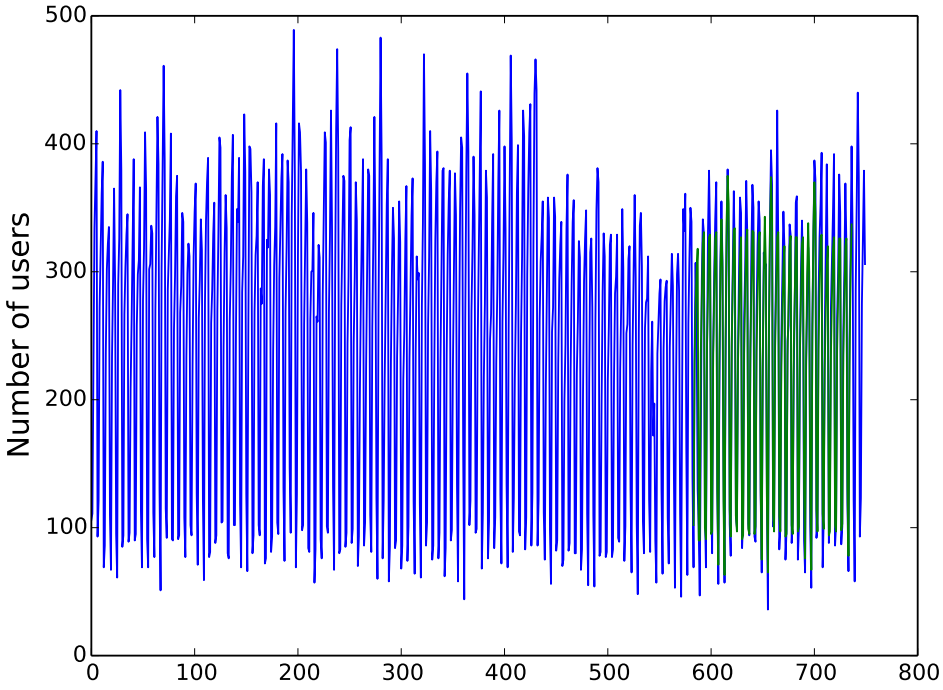

Figure C.12: Arima Allow Drift False - Uniques forecast from 2013-12-31 04:00:00 to 2014-01-25 16:00:00

σ (Real Data)	RMSE	MASE
112.4	30.3	0.0558

Table C.12: Arima Allow Drift False - Error for Uniques forecast from 2013-12-31 04:00:00 to 2014-01-25 16:00:00

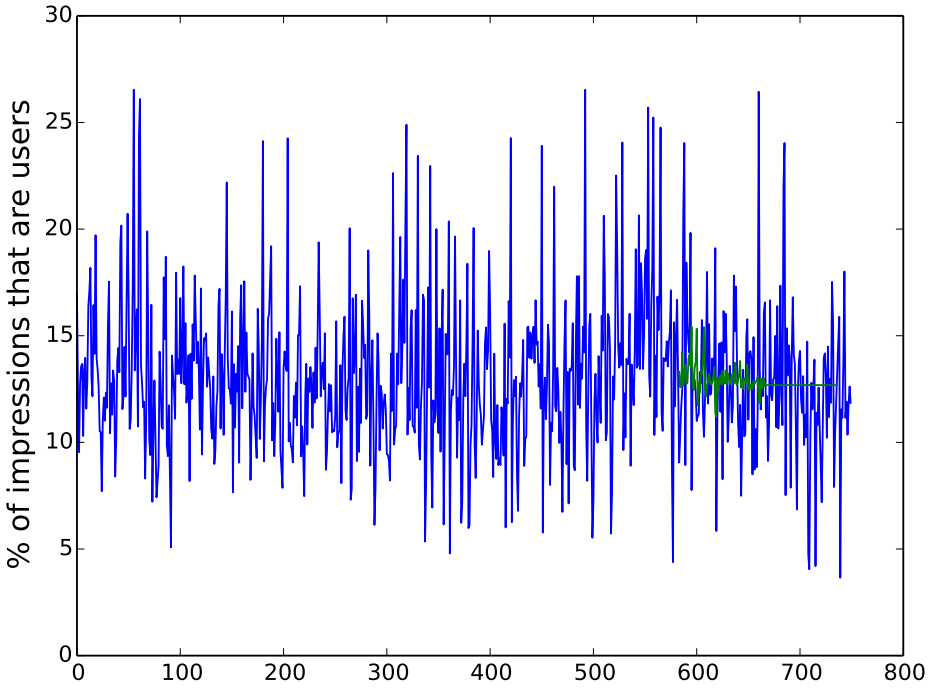

Figure C.13: Arima Allow Drift False - Uniques Percentage forecast from 2013-12-31 04:00:00 to 2014-01-25 16:00:00

σ (Real Data)	RMSE	MASE
3.4	3.49	0.1947

Table C.13: Arima Allow Drift False - Error for Uniques Percentage forecast from 2013-12-31 04:00:00 to 2014-01-25 16:00:00

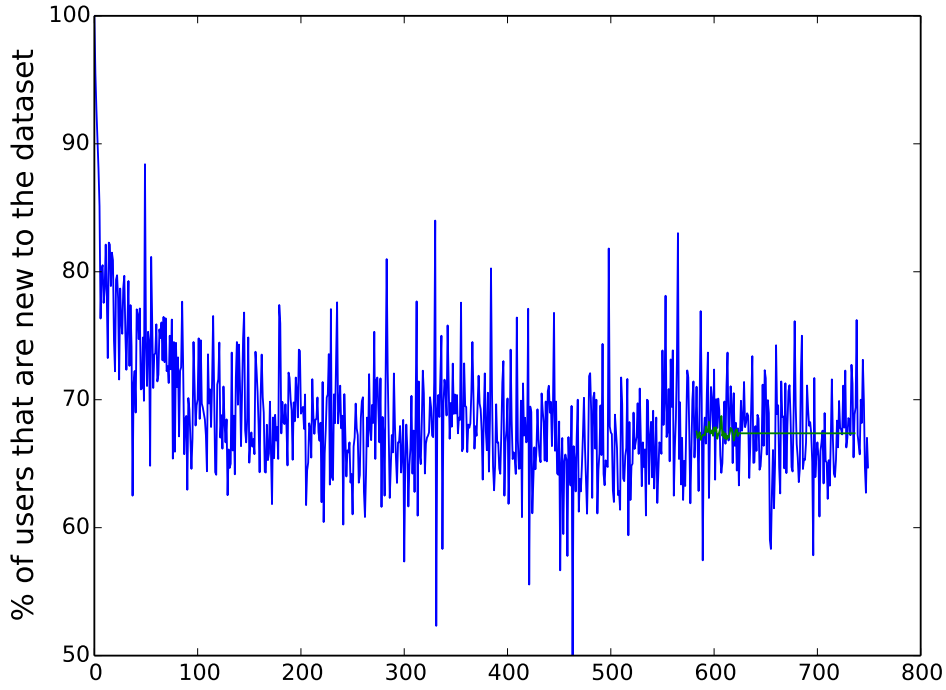

Figure C.14: Arima Allow Drift False - New uniques forecast from 2013-12-31 04:00:00 to 2014-01-25 16:00:00

σ (Real Data)	RMSE	MASE
3.37	3.38	0.1695

Table C.14: Arima Allow Drift False - Error for New Uniques forecast from 2013-12-31 04:00:00 to 2014-01-25 16:00:00

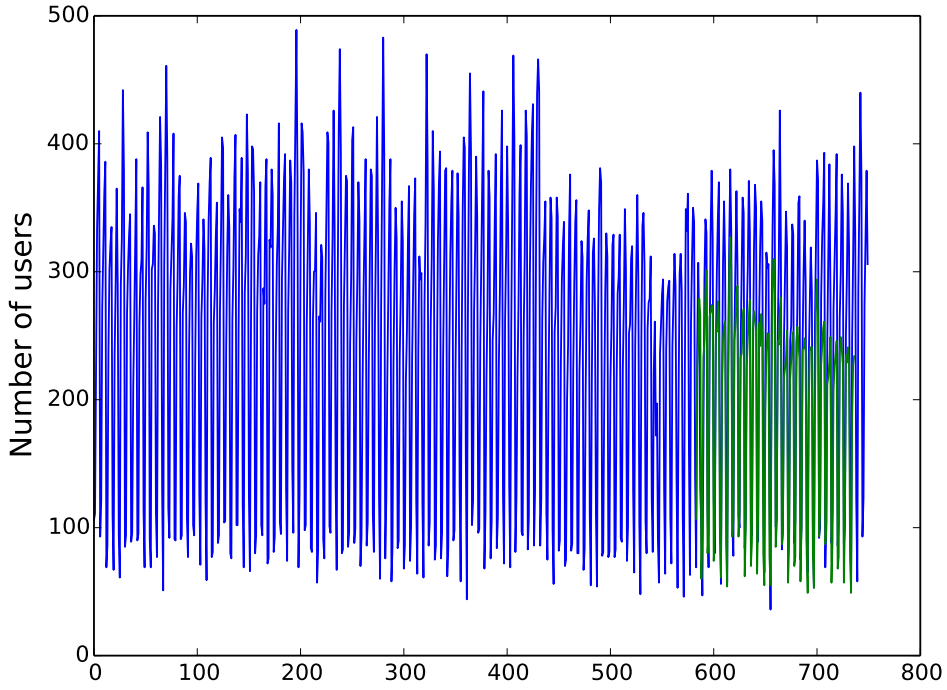

Figure C.15: Arima Allow Drift False - Uniques calculated using percentages forecast from 2013-12-31 04:00:00 to 2014-01-25 16:00:00

σ (Real Data)	RMSE	MASE
112.4	63.21	0.1307

Table C.15: Arima Allow Drift False - Error for Uniques calculated using percentages forecast from 2013-12-31 04:00:00 to 2014-01-25 16:00:00

C.4 Baseline - 6h

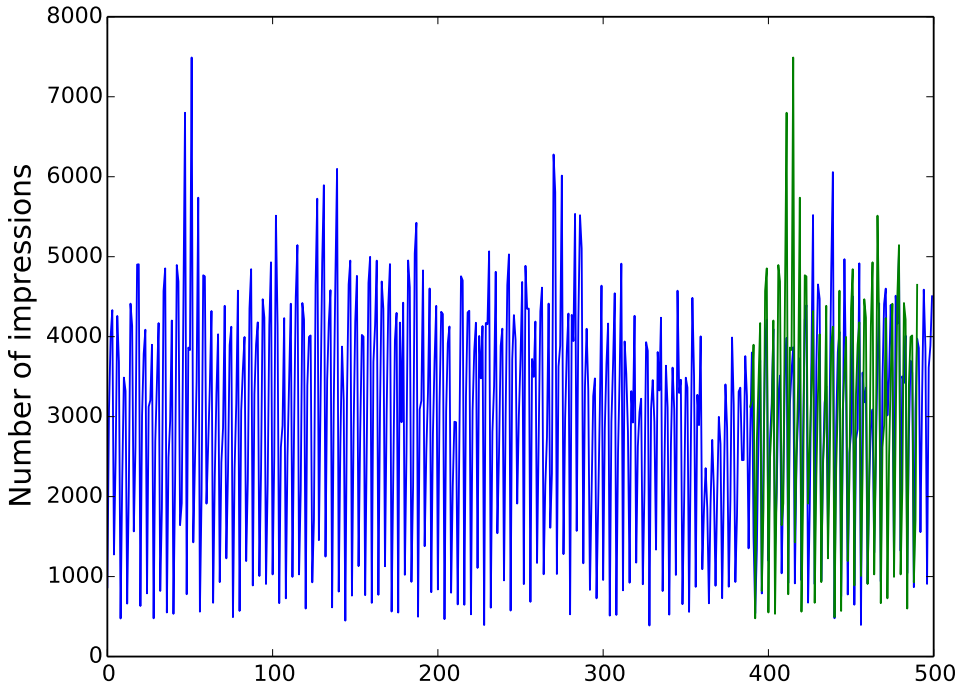

Figure C.16: Baseline - Impressions forecast from 2013-12-31 06:00:00 to 2014-01-25 12:00:00

σ (Real Data)	RMSE	MASE
1334.01	985.86	0.1071

Table C.16: Baseline - Error for Impressions forecast from 2013-12-31 06:00:00 to 2014-01-25 12:00:00

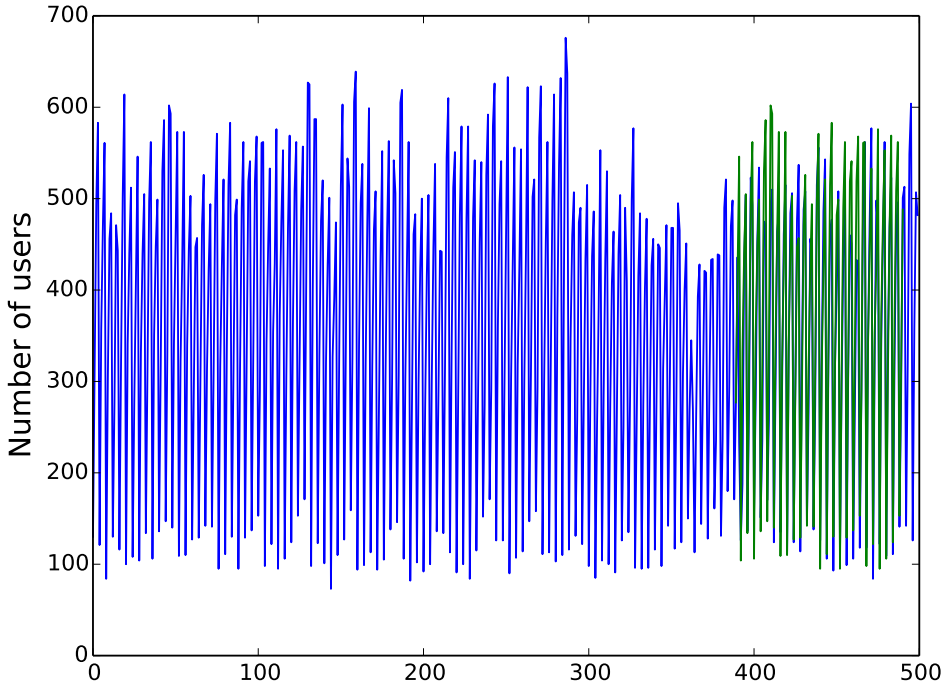

Figure C.17: Baseline - Uniques forecast from 2013-12-31 06:00:00 to 2014-01-25 12:00:00

σ (Real Data)	RMSE	MASE
156.25	57.37	0.0489

Table C.17: Baseline - Error for Uniques forecast from 2013-12-31 06:00:00 to 2014-01-25 12:00:00

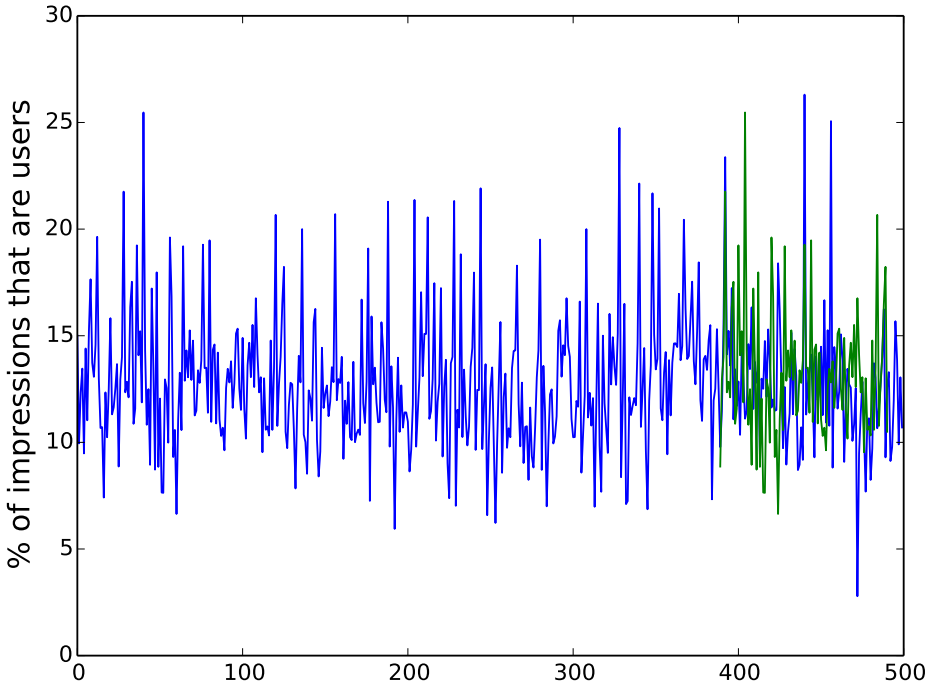

Figure C.18: Baseline - Uniques Percentage forecast from 2013-12-31 06:00:00 to 2014-01-25 12:00:00

σ (Real Data)	RMSE	MASE
3.08	4.34	0.2501

Table C.18: Baseline - Error for Uniques Percentage forecast from 2013-12-31 06:00:00 to 2014-01-25 12:00:00

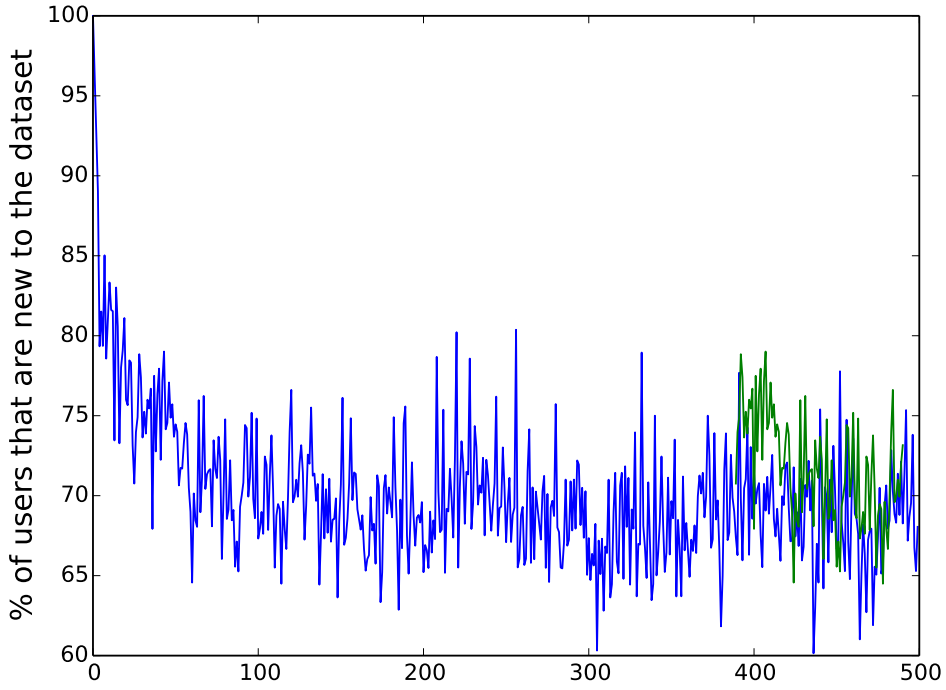

Figure C.19: Baseline - New uniques forecast from 2013-12-31 06:00:00 to 2014-01-25 12:00:00

σ (Real Data)	RMSE	MASE
3.12	4.92	0.3047

Table C.19: Baseline - Error for New Uniques forecast from 2013-12-31 06:00:00 to 2014-01-25 12:00:00

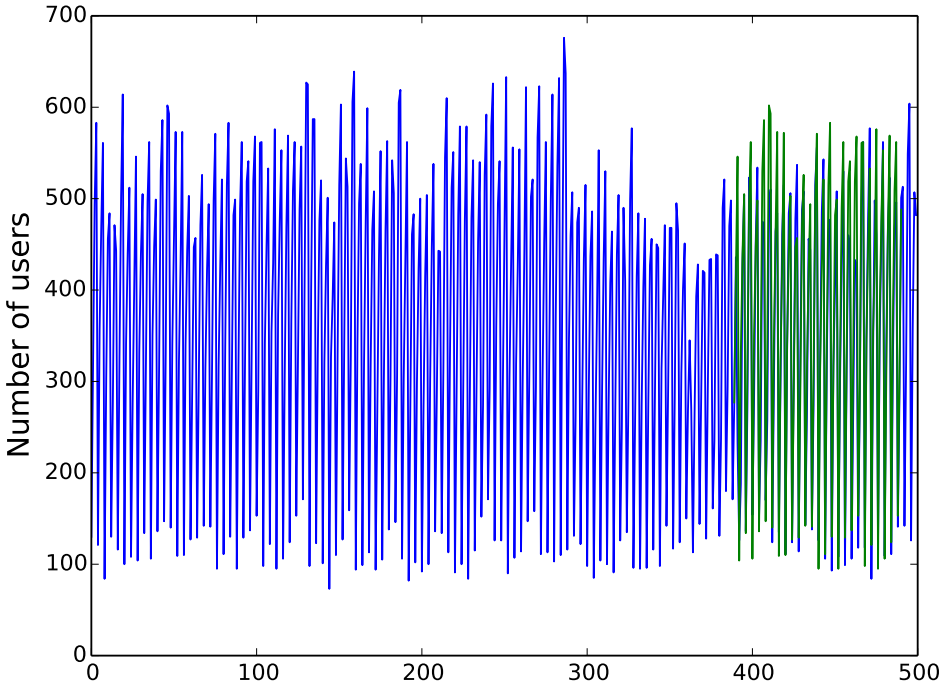

Figure C.20: Baseline - Uniques calculated using percentages forecast from 2013-12-31 06:00:00 to 2014-01-25 12:00:00

σ (Real Data)	RMSE	MASE
156.25	57.37	0.0489

Table C.20: Baseline - Error for Uniques calculated using percentages forecast from 2013-12-31 06:00:00 to 2014-01-25 12:00:00

C.5 Arima Allow Drift True - 6h

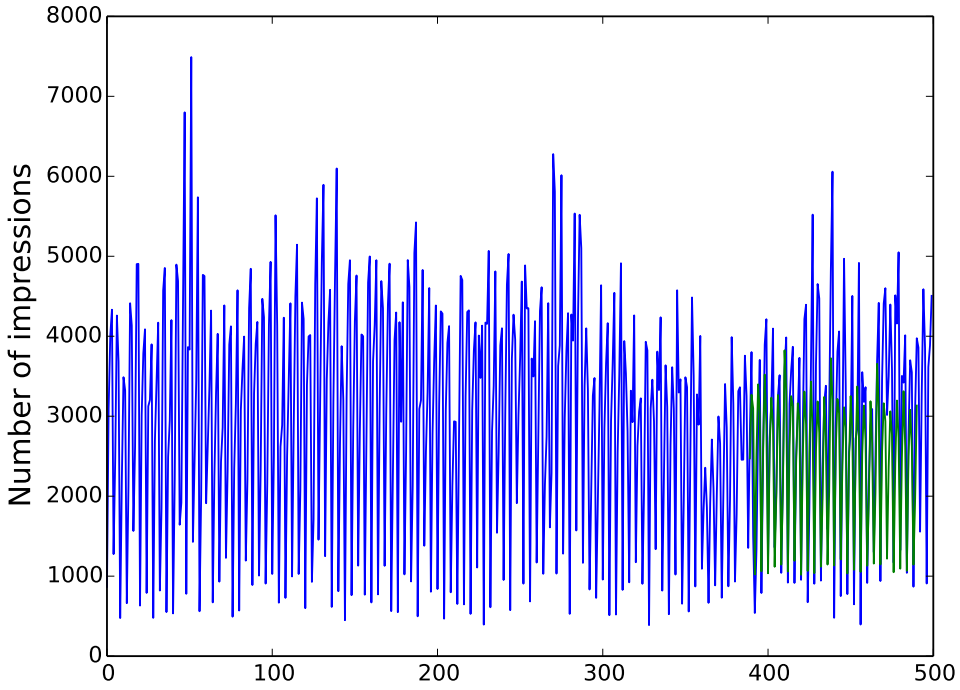

Figure C.21: Arima Allow Drift True - Impressions forecast from 2013-12-31 06:00:00 to 2014-01-25 12:00:00

σ (Real Data)	RMSE	MASE
1334.01	930.13	0.0998

Table C.21: Arima Allow Drift True - Error for Impressions forecast from 2013-12-31 06:00:00 to 2014-01-25 12:00:00

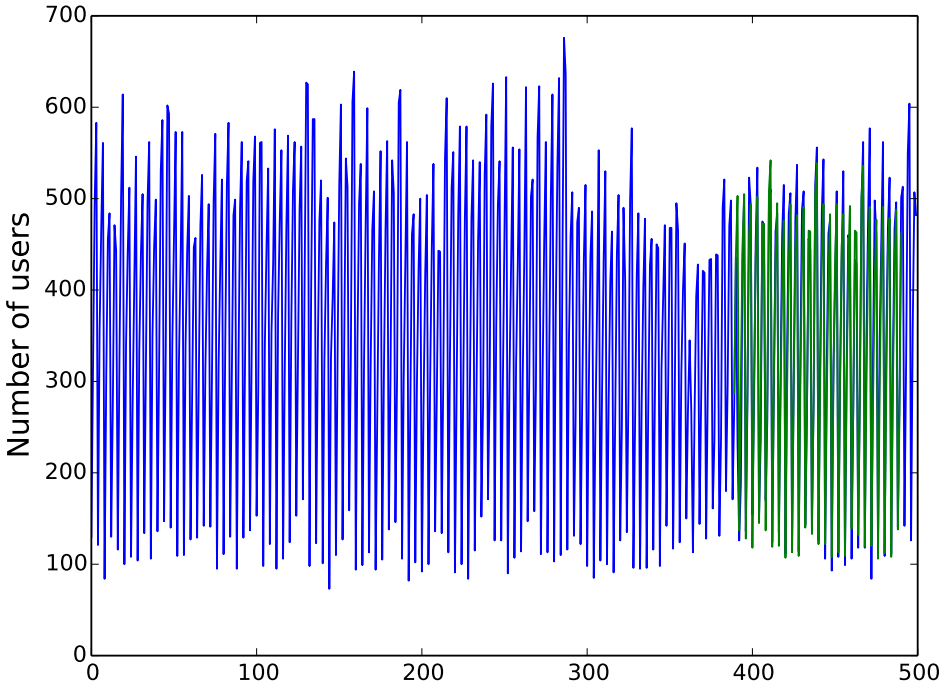

Figure C.22: Arima Allow Drift True - Uniques forecast from 2013-12-31 06:00:00 to 2014-01-25 12:00:00

σ (Real Data)	RMSE	MASE
156.25	42.26	0.0344

Table C.22: Arima Allow Drift True - Error for Uniques forecast from 2013-12-31 06:00:00 to 2014-01-25 12:00:00

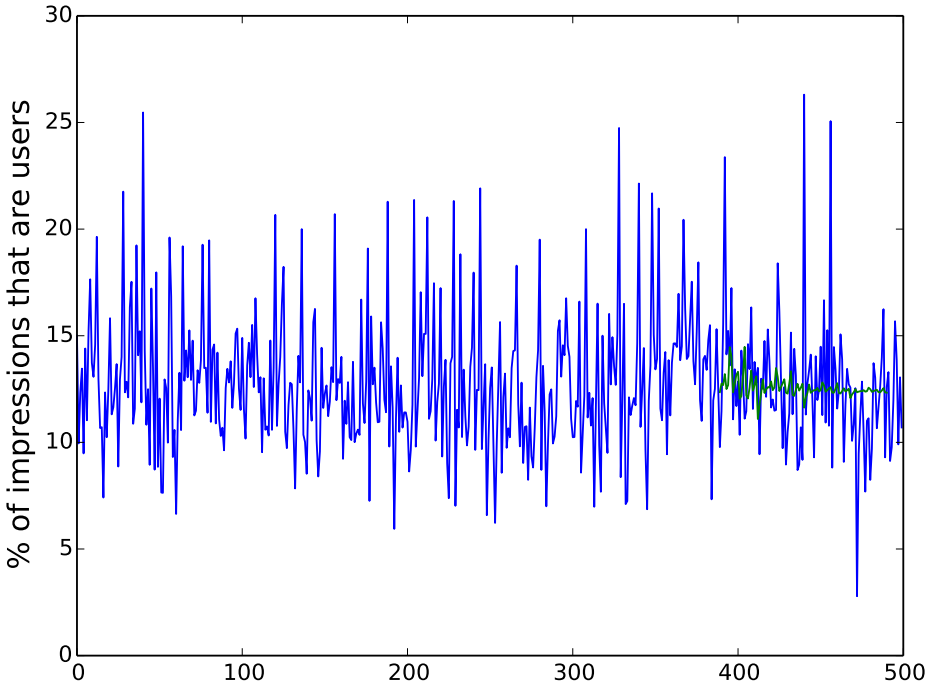

Figure C.23: Arima Allow Drift True - Uniques Percentage forecast from 2013-12-31 06:00:00 to 2014-01-25 12:00:00

σ (Real Data)	RMSE	MASE
3.08	3.13	0.168

Table C.23: Arima Allow Drift True - Error for Uniques Percentage forecast from 2013-12-31 06:00:00 to 2014-01-25 12:00:00

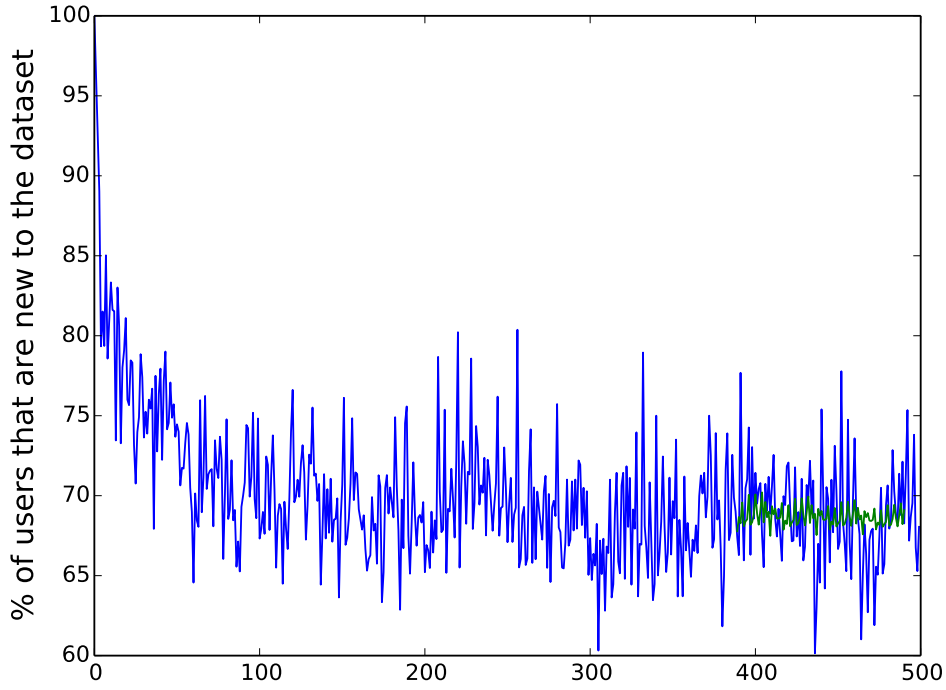

Figure C.24: Arima Allow Drift True - New uniques forecast from 2013-12-31 06:00:00 to 2014-01-25 12:00:00

σ (Real Data)	RMSE	MASE
3.12	2.96	0.1768

Table C.24: Arima Allow Drift True - Error for New Uniques forecast from 2013-12-31 06:00:00 to 2014-01-25 12:00:00

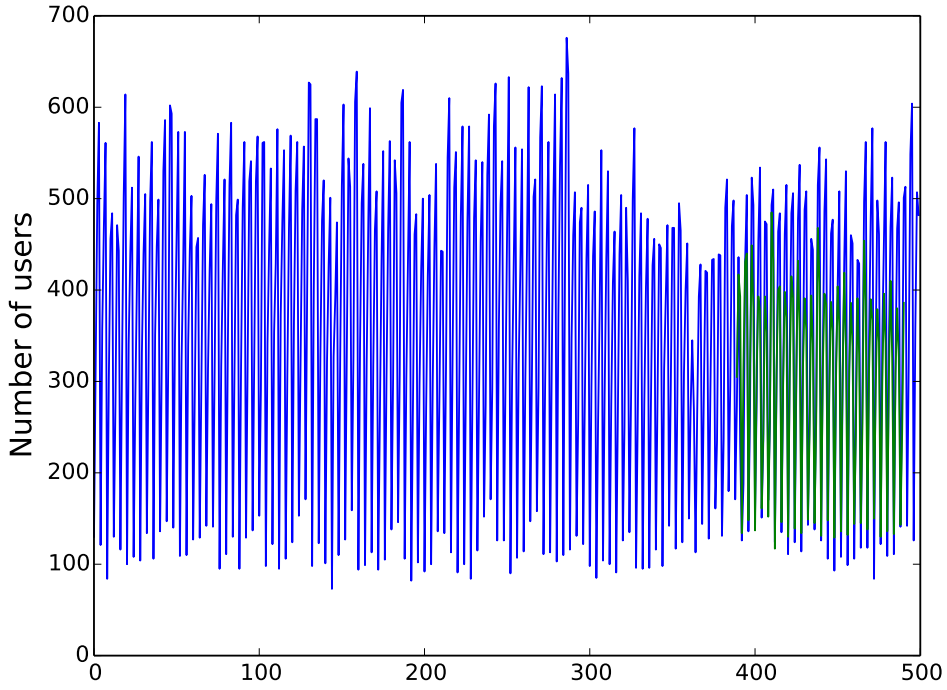

Figure C.25: Arima Allow Drift True - Uniques calculated using percentages forecast from 2013-12-31 06:00:00 to 2014-01-25 12:00:00

σ (Real Data)	RMSE	MASE
156.25	83.52	0.0772

Table C.25: Arima Allow Drift True - Error for Uniques calculated using percentages forecast from 2013-12-31 06:00:00 to 2014-01-25 12:00:00

C.6 Arima Allow Drift False - 6h

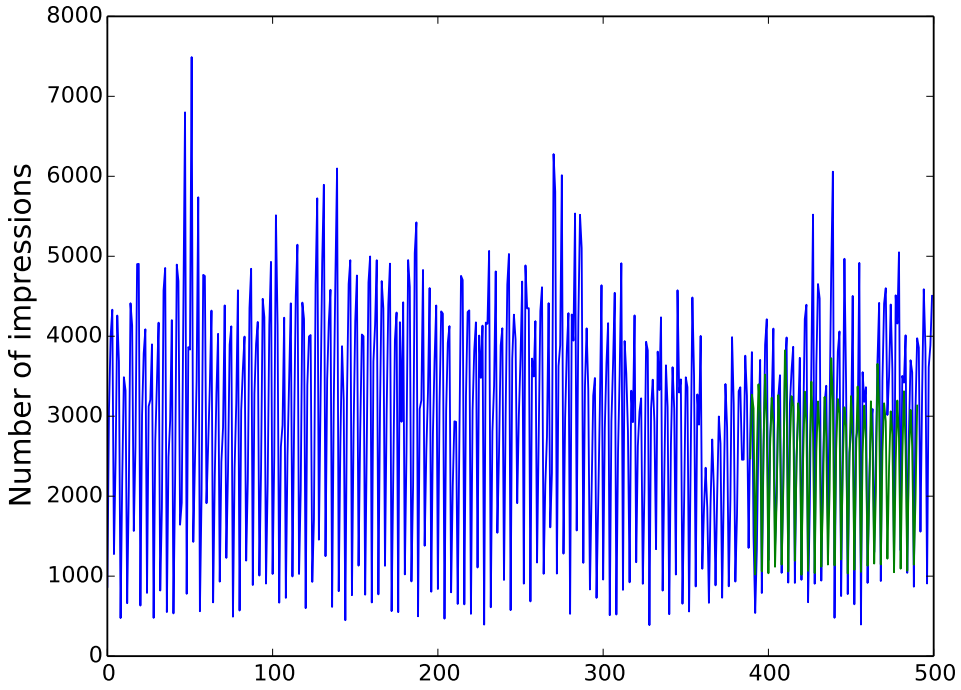

Figure C.26: Arima Allow Drift False - Impressions forecast from 2013-12-31 06:00:00 to 2014-01-25 12:00:00

σ (Real Data)	RMSE	MASE
1334.01	930.13	0.0998

Table C.26: Arima Allow Drift False - Error for Impressions forecast from 2013-12-31 06:00:00 to 2014-01-25 12:00:00

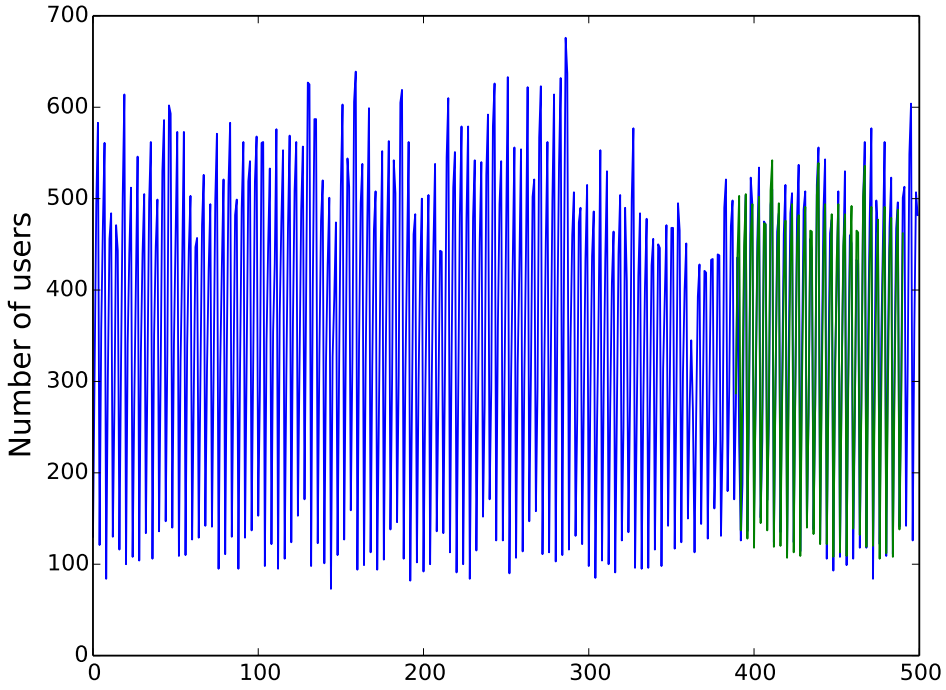

Figure C.27: Arima Allow Drift False - Uniques forecast from 2013-12-31 06:00:00 to 2014-01-25 12:00:00

σ (Real Data)	RMSE	MASE
156.25	42.26	0.0344

Table C.27: Arima Allow Drift False - Error for Uniques forecast from 2013-12-31 06:00:00 to 2014-01-25 12:00:00

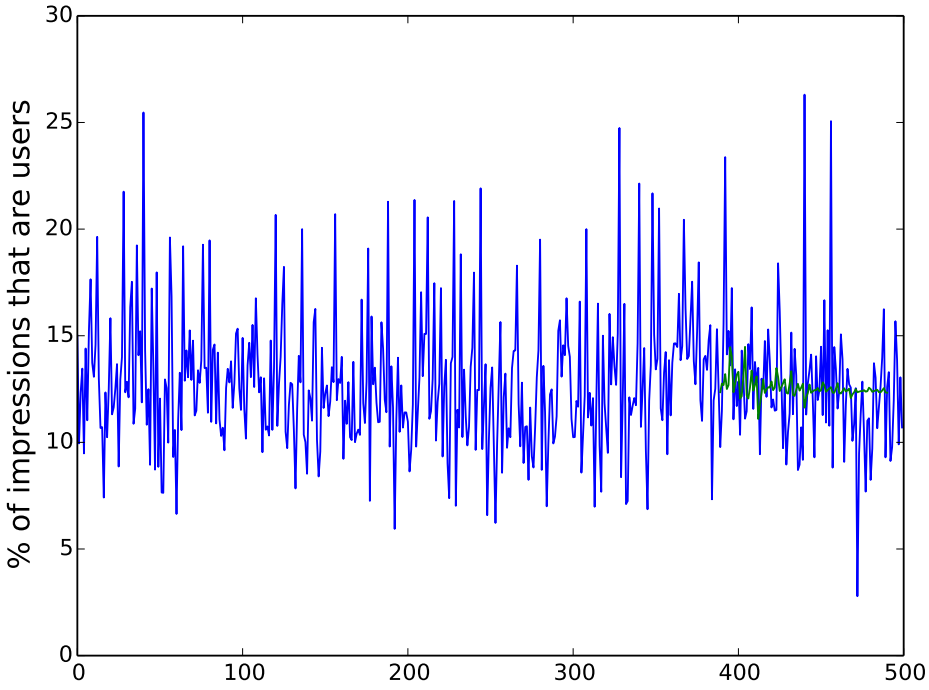

Figure C.28: Arima Allow Drift False - Uniques Percentage forecast from 2013-12-31 06:00:00 to 2014-01-25 12:00:00

σ (Real Data)	RMSE	MASE
3.08	3.13	0.168

Table C.28: Arima Allow Drift False - Error for Uniques Percentage forecast from 2013-12-31 06:00:00 to 2014-01-25 12:00:00

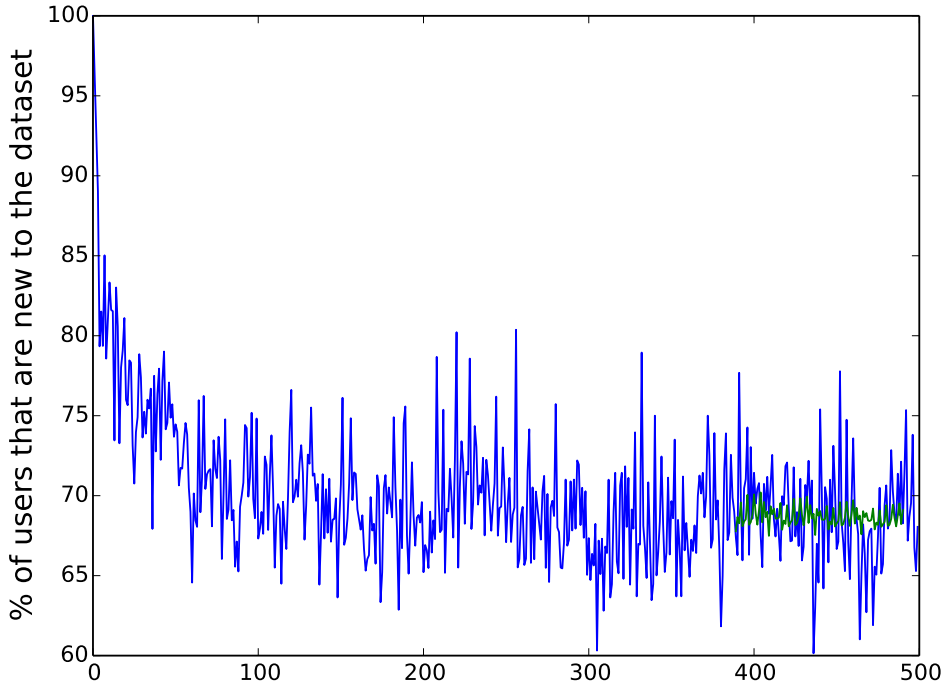

Figure C.29: Arima Allow Drift False - New uniques forecast from 2013-12-31 06:00:00 to 2014-01-25 12:00:00

σ (Real Data)	RMSE	MASE
3.12	2.96	0.1768

Table C.29: Arima Allow Drift False - Error for New Uniques forecast from 2013-12-31 06:00:00 to 2014-01-25 12:00:00

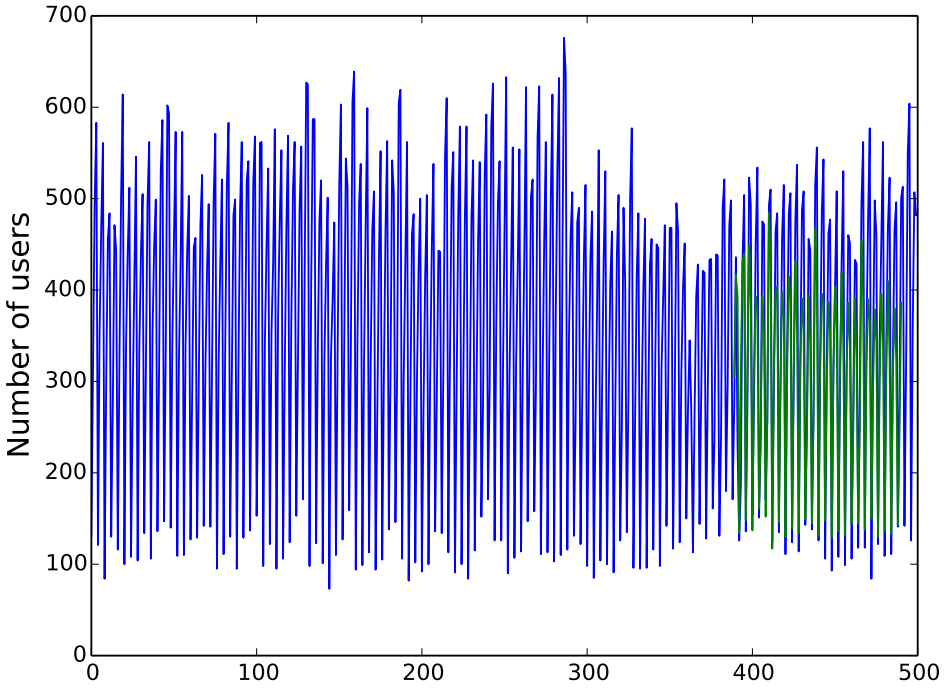

Figure C.30: Arima Allow Drift False - Uniques calculated using percentages forecast from 2013-12-31 06:00:00 to 2014-01-25 12:00:00

σ (Real Data)	RMSE	MASE
156.25	83.52	0.0772

Table C.30: Arima Allow Drift False - Error for Uniques calculated using percentages forecast from 2013-12-31 06:00:00 to 2014-01-25 12:00:00

C.7 Baseline - 8h

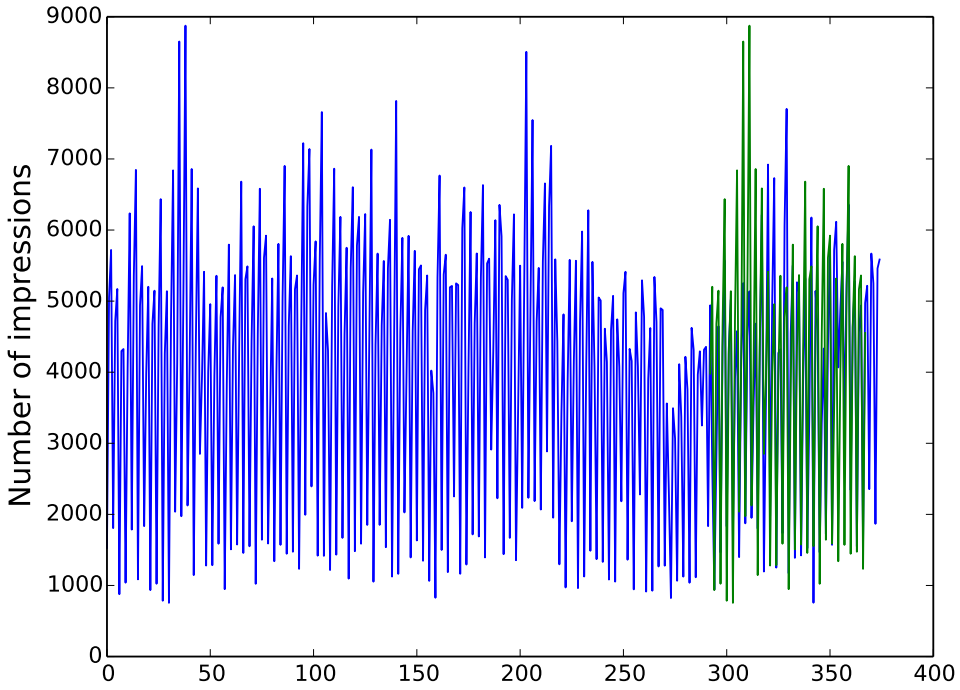

Figure C.31: Baseline - Impressions forecast from 2013-12-31 08:00:00 to 2014-01-25 08:00:00

σ (Real Data)	RMSE	MASE
1738.75	1197.45	0.0778

Table C.31: Baseline - Error for Impressions forecast from 2013-12-31 08:00:00 to 2014-01-25 08:00:00

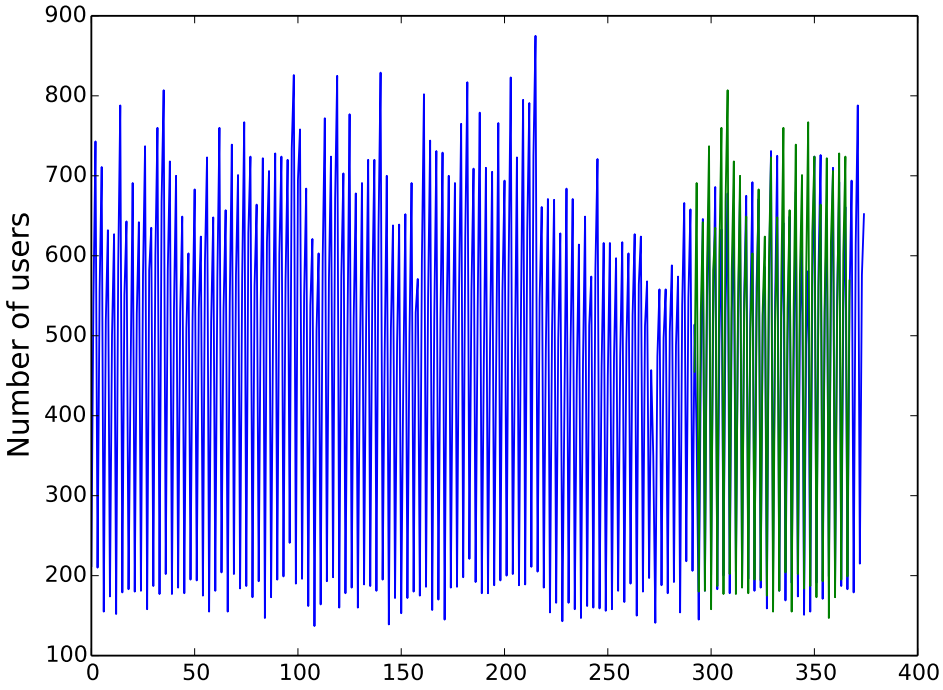

Figure C.32: Baseline - Uniques forecast from 2013-12-31 08:00:00 to 2014-01-25 08:00:00

σ (Real Data)	RMSE	MASE
206.66	67.28	0.0355

Table C.32: Baseline - Error for Uniques forecast from 2013-12-31 08:00:00 to 2014-01-25 08:00:00

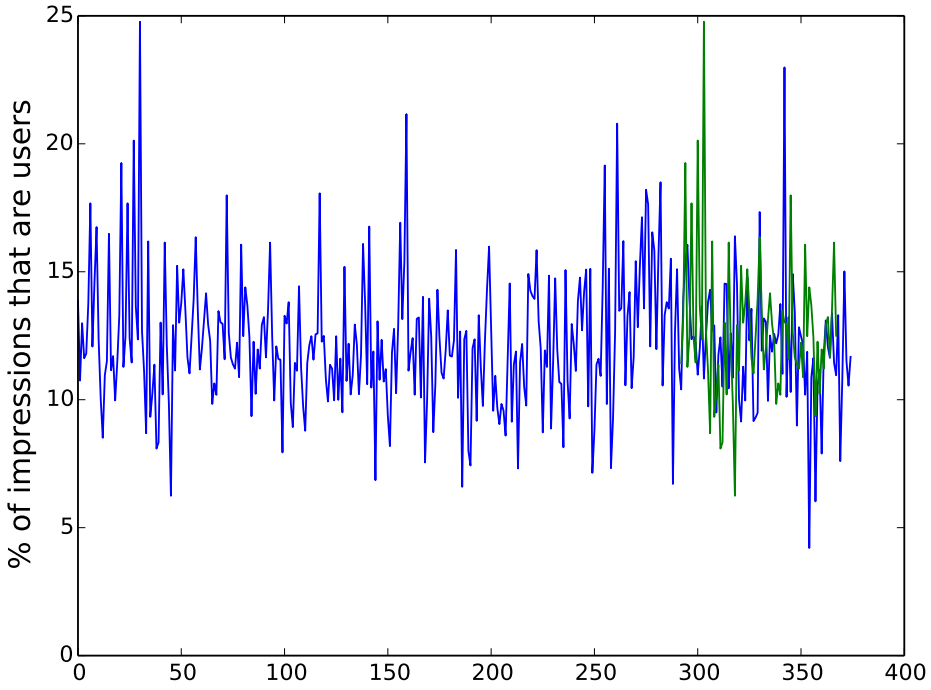

Figure C.33: Baseline - Uniques Percentage forecast from 2013-12-31 08:00:00 to 2014-01-25 08:00:00

σ (Real Data)	RMSE	MASE
2.46	3.87	0.2611

Table C.33: Baseline - Error for Uniques Percentage forecast from 2013-12-31 08:00:00 to 2014-01-25 08:00:00

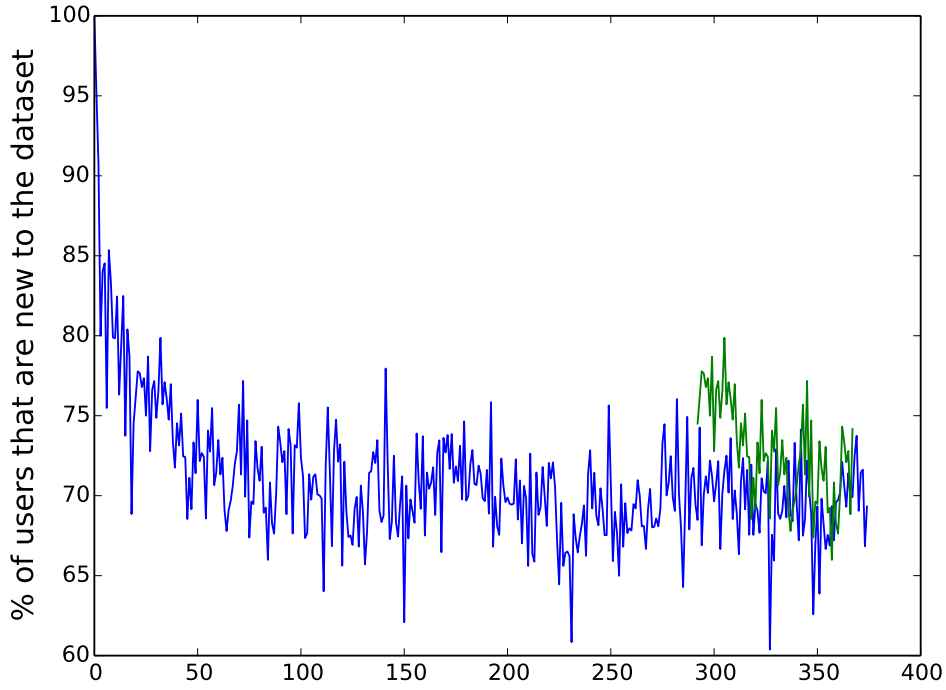

Figure C.34: Baseline - New uniques forecast from 2013-12-31 08:00:00 to 2014-01-25 08:00:00

σ (Real Data)	RMSE	MASE
2.43	4.61	0.3589

Table C.34: Baseline - Error for New Uniques forecast from 2013-12-31 08:00:00 to 2014-01-25 08:00:00

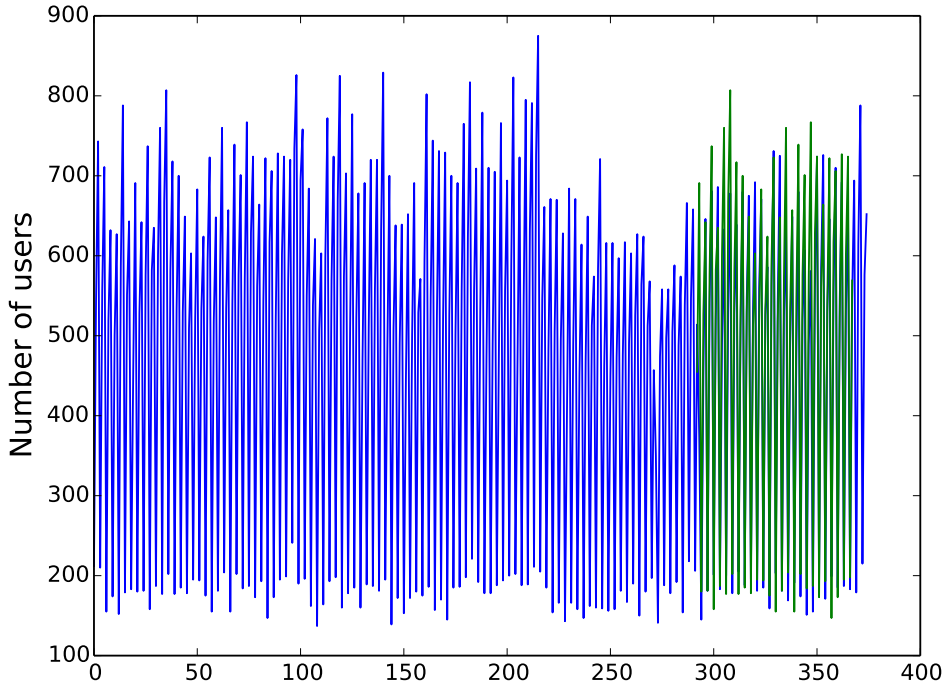

Figure C.35: Baseline - Uniques calculated using percentages forecast from 2013-12-31 08:00:00 to 2014-01-25 08:00:00

σ (Real Data)	RMSE	MASE
206.66	67.24	0.0355

Table C.35: Baseline - Error for Uniques calculated using percentages forecast from 2013-12-31 08:00:00 to 2014-01-25 08:00:00

C.8 Arima Allow Drift True - 8h

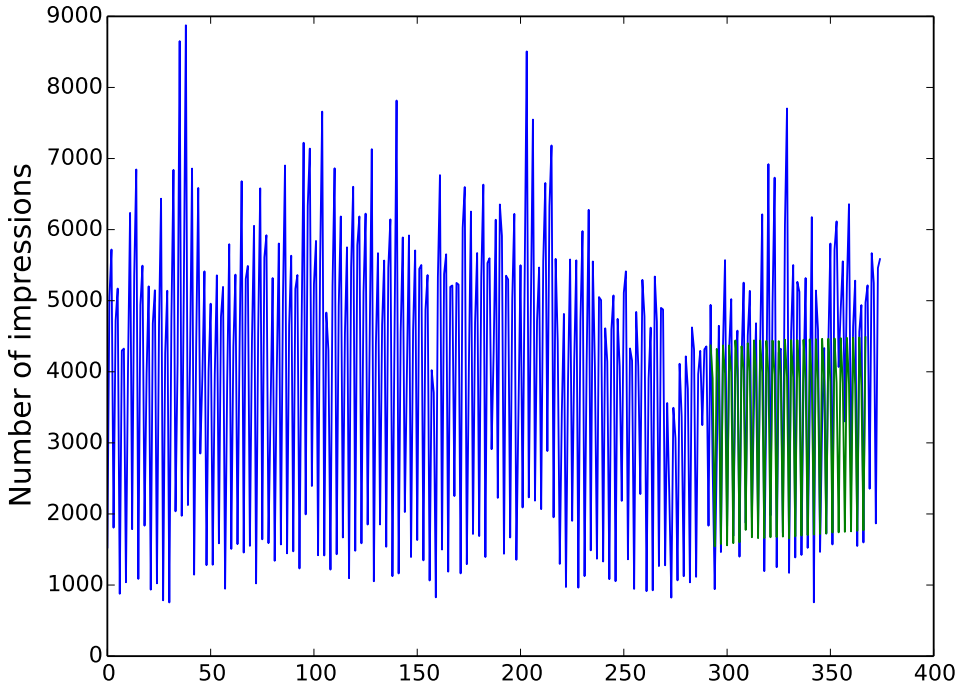

Figure C.36: Arima Allow Drift True - Impressions forecast from 2013-12-31 08:00:00 to 2014-01-25 08:00:00

σ (Real Data)	RMSE	MASE
1738.75	1165.13	0.0763

Table C.36: Arima Allow Drift True - Error for Impressions forecast from 2013-12-31 08:00:00 to 2014-01-25 08:00:00

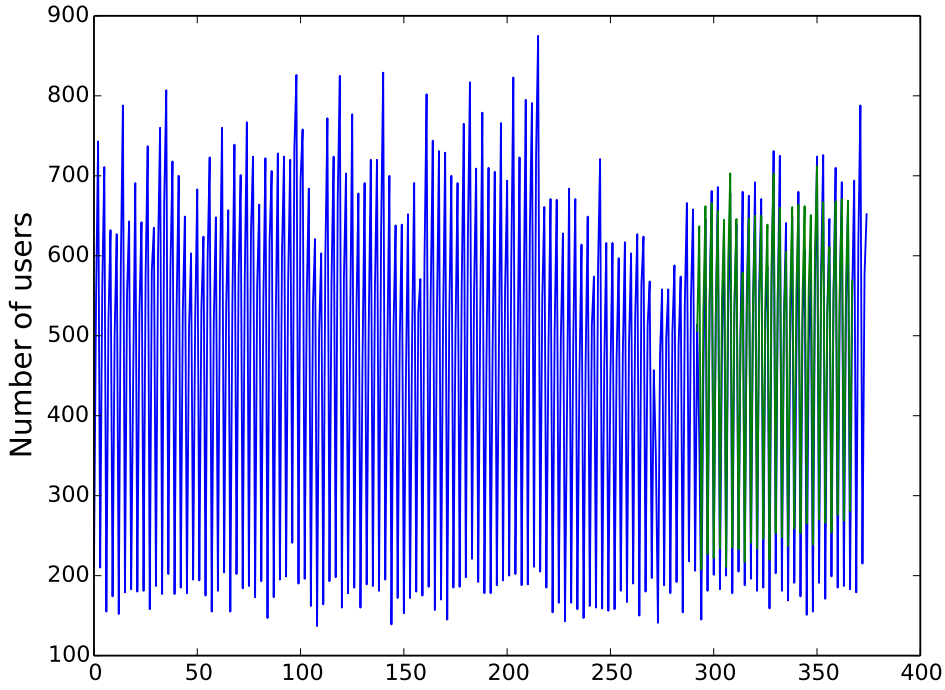

Figure C.37: Arima Allow Drift True - Uniques forecast from 2013-12-31 08:00:00 to 2014-01-25 08:00:00

σ (Real Data)	RMSE	MASE
206.66	58.12	0.0333

Table C.37: Arima Allow Drift True - Error for Uniques forecast from 2013-12-31 08:00:00 to 2014-01-25 08:00:00

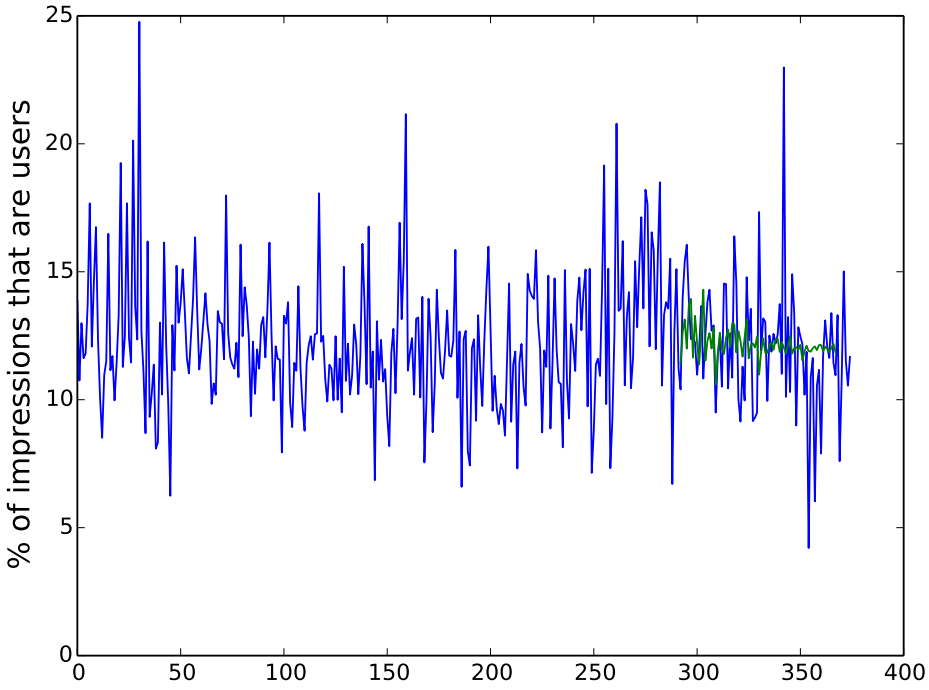

Figure C.38: Arima Allow Drift True - Uniques Percentage forecast from 2013-12-31 08:00:00 to 2014-01-25 08:00:00

σ (Real Data)	RMSE	MASE
2.46	2.5	0.1643

Table C.38: Arima Allow Drift True - Error for Uniques Percentage forecast from 2013-12-31 08:00:00 to 2014-01-25 08:00:00

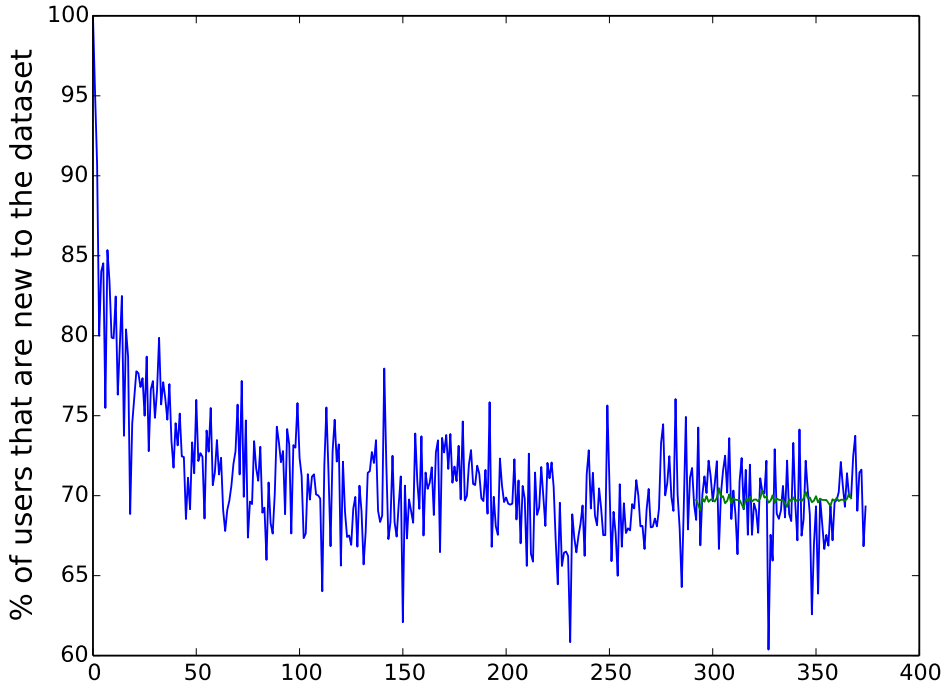

Figure C.39: Arima Allow Drift True - New uniques forecast from 2013-12-31 08:00:00 to 2014-01-25 08:00:00

σ (Real Data)	RMSE	MASE
2.43	2.42	0.1645

Table C.39: Arima Allow Drift True - Error for New Uniques forecast from 2013-12-31 08:00:00 to 2014-01-25 08:00:00

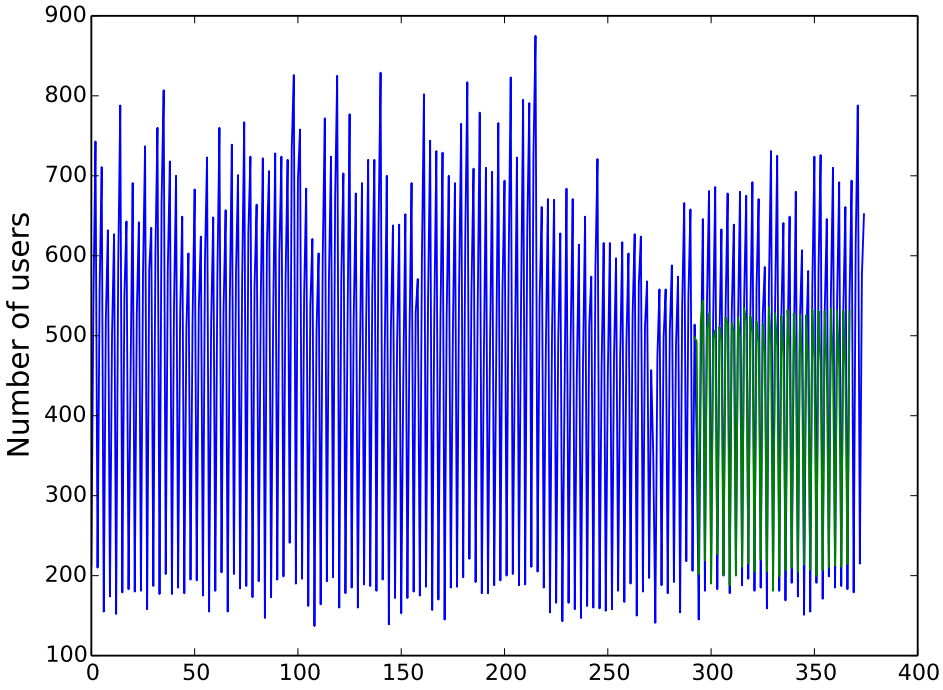

Figure C.40: Arima Allow Drift True - Uniques calculated using percentages forecast from 2013-12-31 08:00:00 to 2014-01-25 08:00:00

σ (Real Data)	RMSE	MASE
206.66	108.76	0.057

Table C.40: Arima Allow Drift True - Error for Uniques calculated using percentages forecast from 2013-12-31 08:00:00 to 2014-01-25 08:00:00

C.9 Arima Allow Drift False - 8h

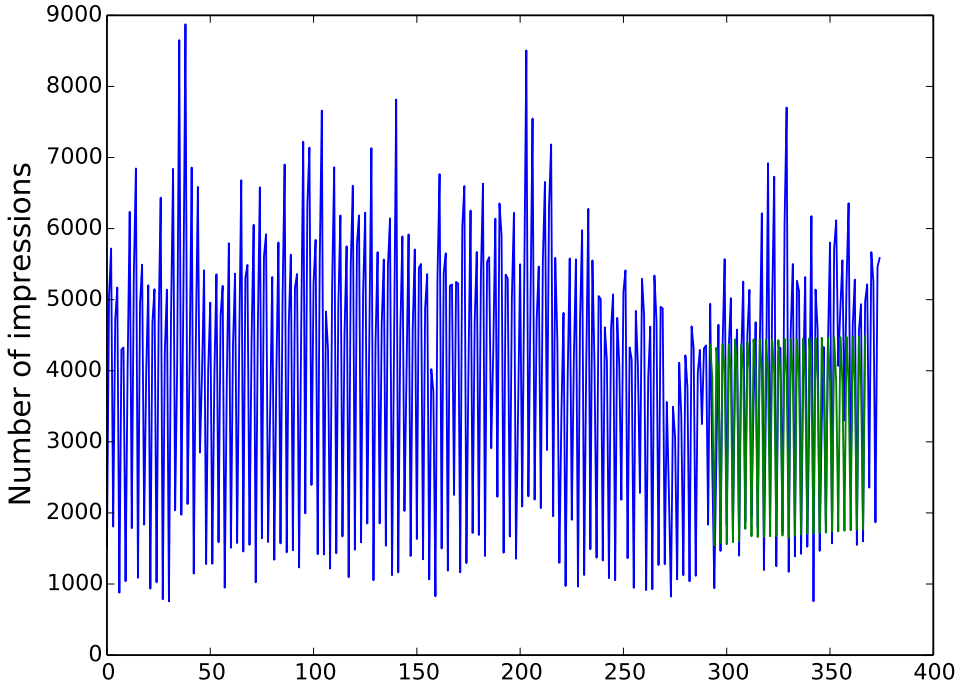

Figure C.41: Arima Allow Drift False - Impressions forecast from 2013-12-31 08:00:00 to 2014-01-25 08:00:00

σ (Real Data)	RMSE	MASE
1738.75	1165.13	0.0763

Table C.41: Arima Allow Drift False - Error for Impressions forecast from 2013-12-31 08:00:00 to 2014-01-25 08:00:00

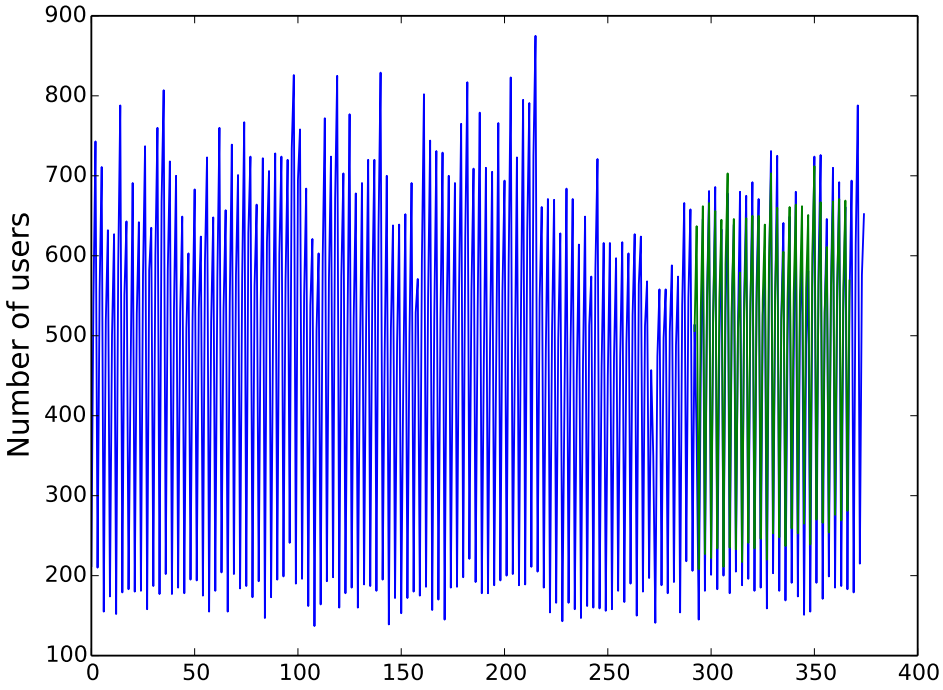

Figure C.42: Arima Allow Drift False - Uniques forecast from 2013-12-31 08:00:00 to 2014-01-25 08:00:00

σ (Real Data)	RMSE	MASE
206.66	58.12	0.0333

Table C.42: Arima Allow Drift False - Error for Uniques forecast from 2013-12-31 08:00:00 to 2014-01-25 08:00:00

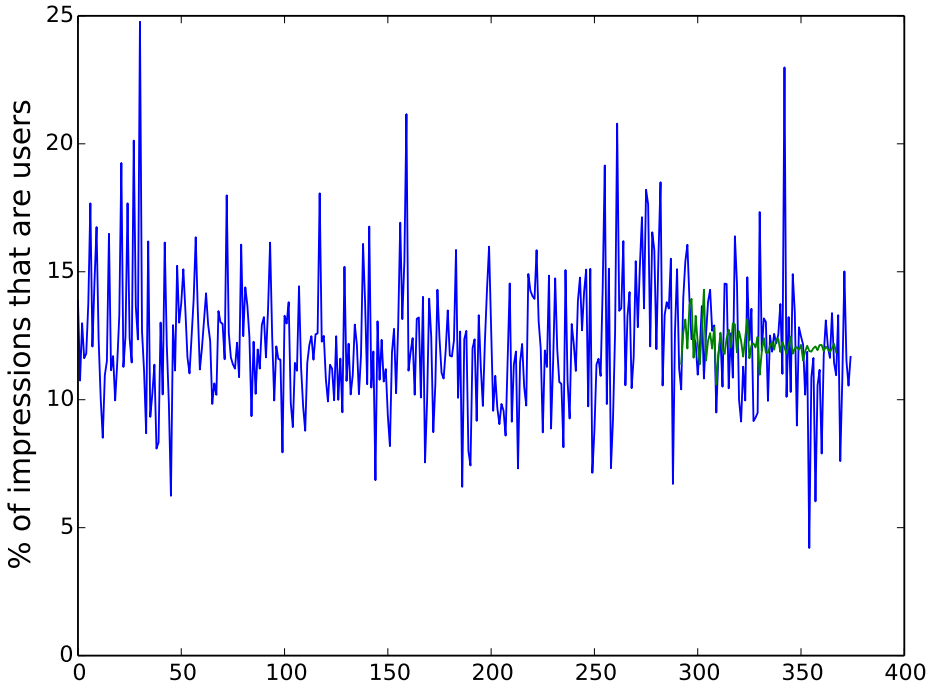

Figure C.43: Arima Allow Drift False - Uniques Percentage forecast from 2013-12-31 08:00:00 to 2014-01-25 08:00:00

σ (Real Data)	RMSE	MASE
2.46	2.5	0.1643

Table C.43: Arima Allow Drift False - Error for Uniques Percentage forecast from 2013-12-31 08:00:00 to 2014-01-25 08:00:00

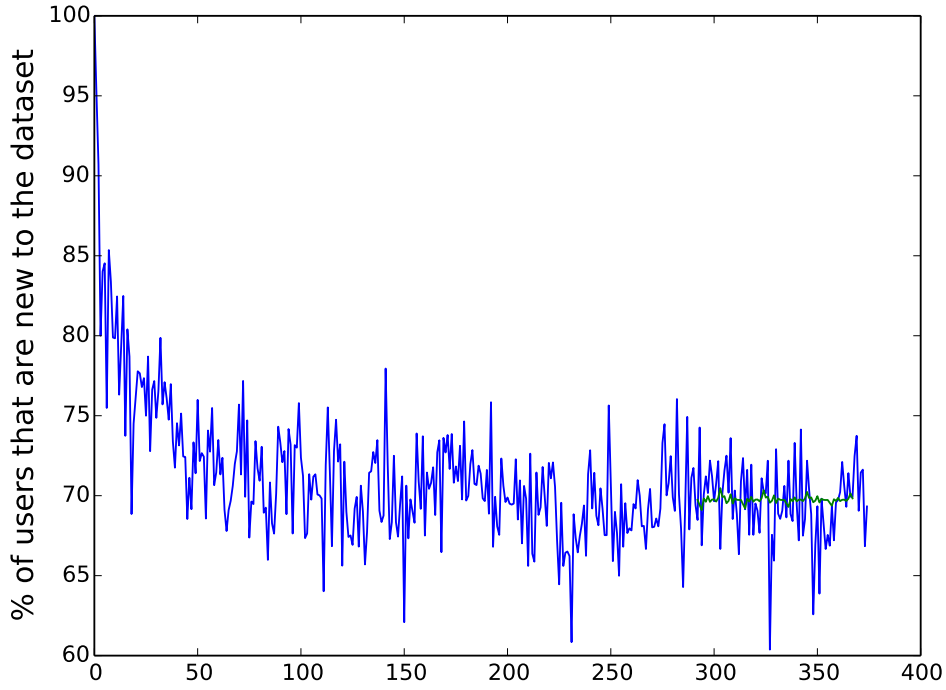

Figure C.44: Arima Allow Drift False - New uniques forecast from 2013-12-31 08:00:00 to 2014-01-25 08:00:00

σ (Real Data)	RMSE	MASE
2.43	2.42	0.1645

Table C.44: Arima Allow Drift False - Error for New Uniques forecast from 2013-12-31 08:00:00 to 2014-01-25 08:00:00

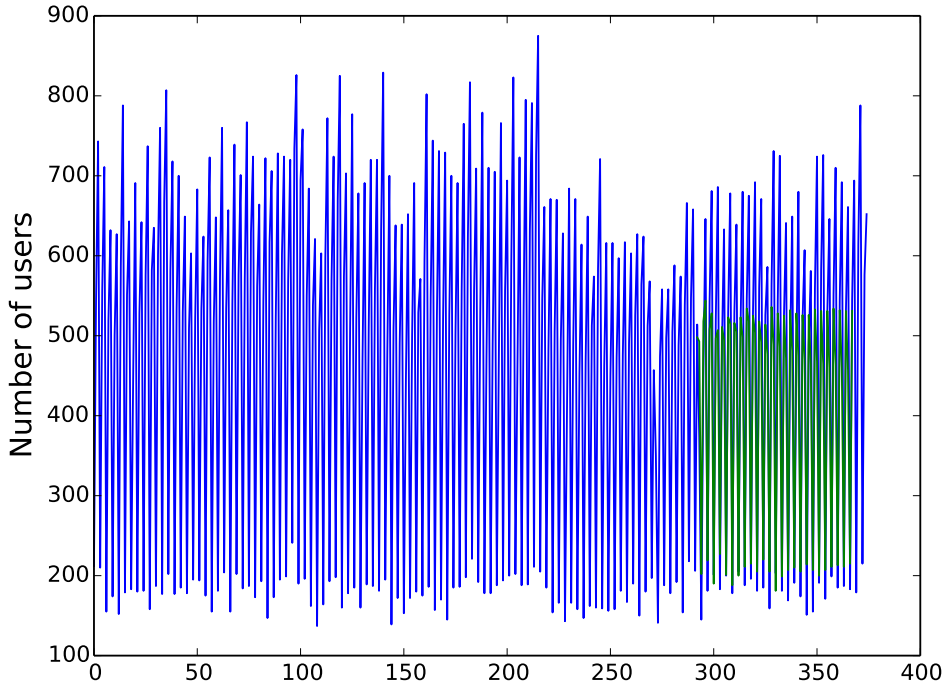

Figure C.45: Arima Allow Drift False - Uniques calculated using percentages forecast from 2013-12-31 08:00:00 to 2014-01-25 08:00:00

σ (Real Data)	RMSE	MASE
206.66	108.76	0.057

Table C.45: Arima Allow Drift False - Error for Uniques calculated using percentages forecast from 2013-12-31 08:00:00 to 2014-01-25 08:00:00

C.10 Baseline - 12h

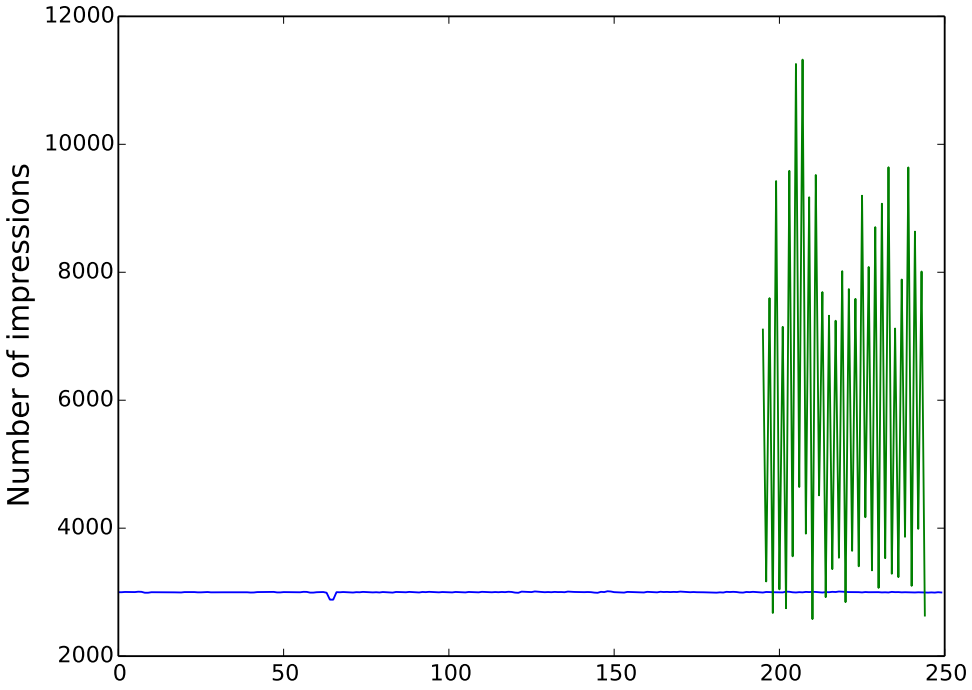

Figure C.46: Baseline - Impressions forecast from 2013-12-31 12:00:00 to 2014-01-25 00:00:00

σ (Real Data)	RMSE	MASE
3.76	4041.36	166.6978

Table C.46: Baseline - Error for Impressions forecast from 2013-12-31 12:00:00 to 2014-01-25 00:00:00

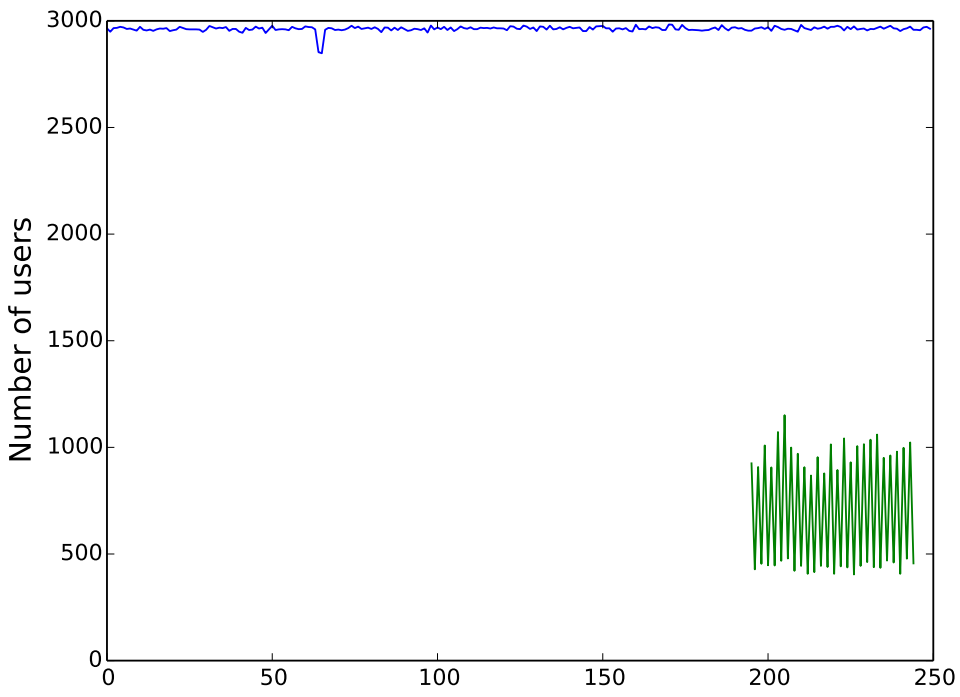

Figure C.47: Baseline - Uniques forecast from 2013-12-31 12:00:00 to 2014-01-25 00:00:00

σ (Real Data)	RMSE	MASE
7.13	2271.66	61.2845

Table C.47: Baseline - Error for Uniques forecast from 2013-12-31 12:00:00 to 2014-01-25 00:00:00

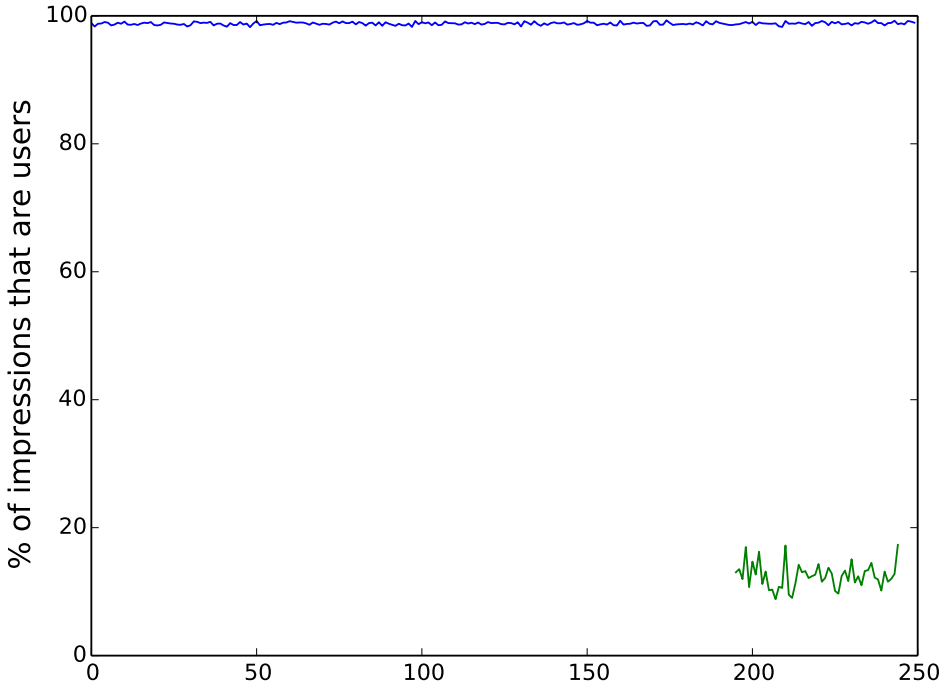

Figure C.48: Baseline - Uniques Percentage forecast from 2013-12-31 12:00:00 to 2014-01-25 00:00:00

σ (Real Data)	RMSE	MASE
0.22	86.39	88.948

Table C.48: Baseline - Error for Uniques Percentage forecast from 2013-12-31 12:00:00 to 2014-01-25 00:00:00

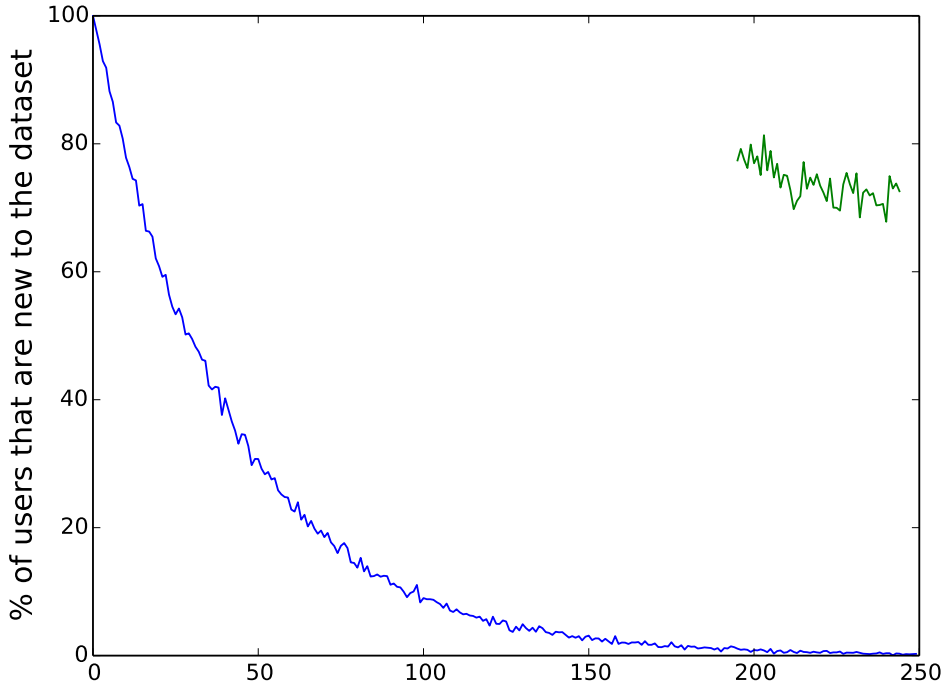

Figure C.49: Baseline - New uniques forecast from 2013-12-31 12:00:00 to 2014-01-25 00:00:00

σ (Real Data)	RMSE	MASE
0.25	73.37	21.5586

Table C.49: Baseline - Error for New Uniques forecast from 2013-12-31 12:00:00 to 2014-01-25 00:00:00

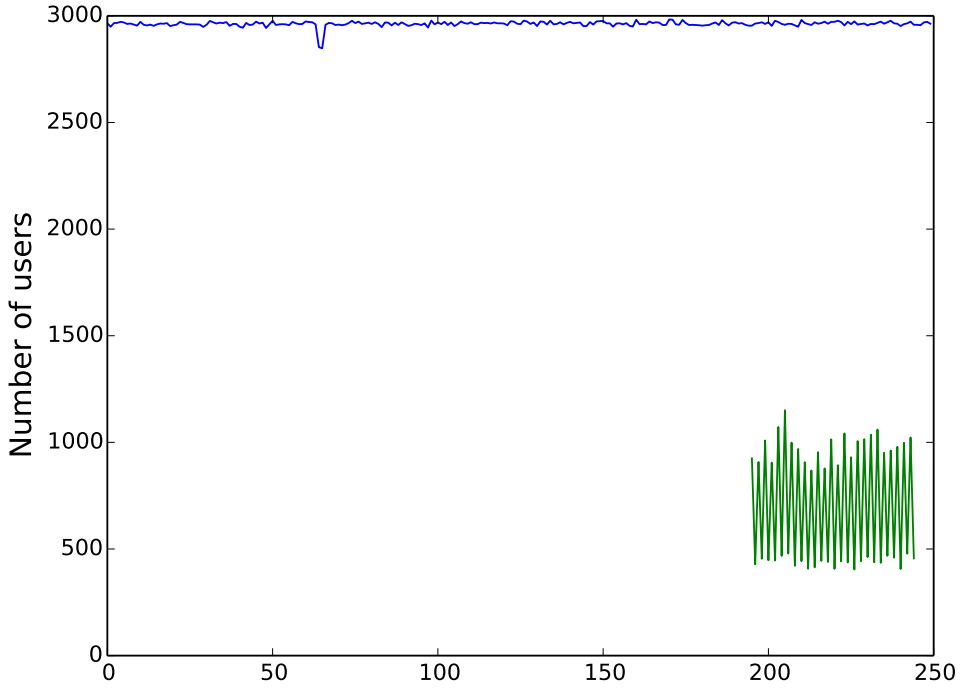

Figure C.50: Baseline - Uniques calculated using percentages forecast from 2013-12-31 12:00:00 to 2014-01-25 00:00:00

σ (Real Data)	RMSE	MASE
7.13	2271.87	61.29

Table C.50: Baseline - Error for Uniques calculated using percentages forecast from 2013-12-31 12:00:00 to 2014-01-25 00:00:00

C.11 Arima Allow Drift True - 12h

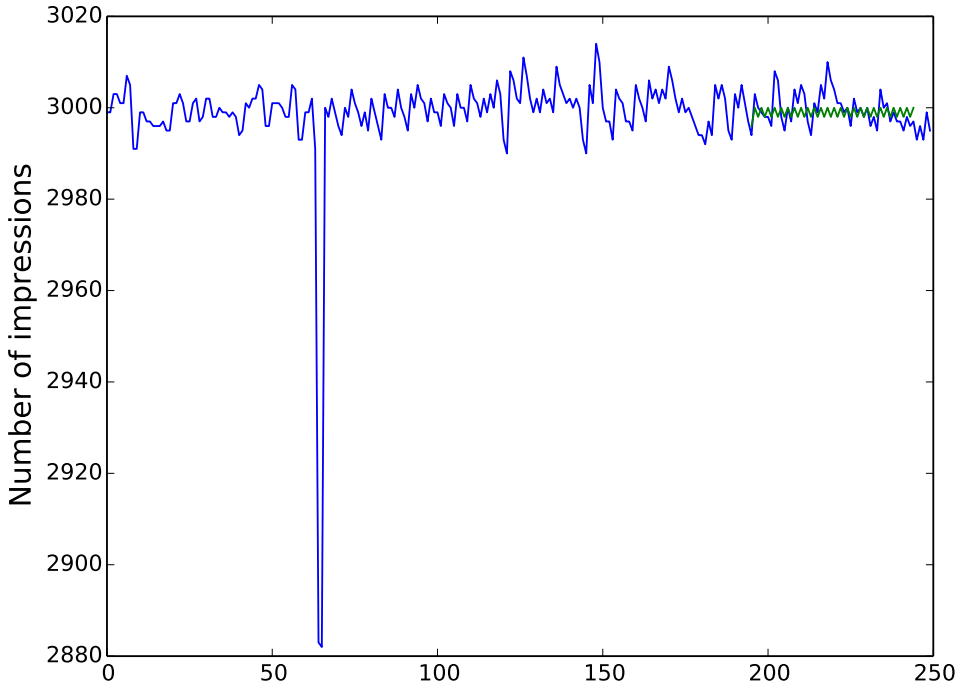

Figure C.51: Arima Allow Drift True - Impressions forecast from 2013-12-31 12:00:00 to 2014-01-25 00:00:00

σ (Real Data)	RMSE	MASE
3.76	3.43	0.1439

Table C.51: Arima Allow Drift True - Error for Impressions forecast from 2013-12-31 12:00:00 to 2014-01-25 00:00:00

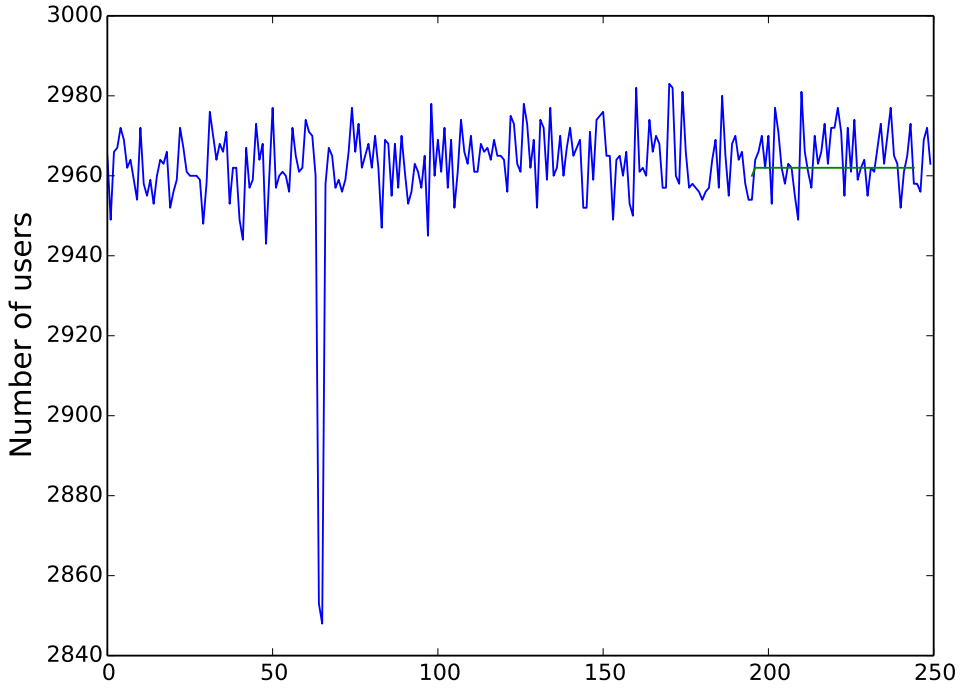

Figure C.52: Arima Allow Drift True - Uniques forecast from 2013-12-31 12:00:00 to 2014-01-25 00:00:00

σ (Real Data)	RMSE	MASE
7.13	7.7	0.1631

Table C.52: Arima Allow Drift True - Error for Uniques forecast from 2013-12-31 12:00:00 to 2014-01-25 00:00:00

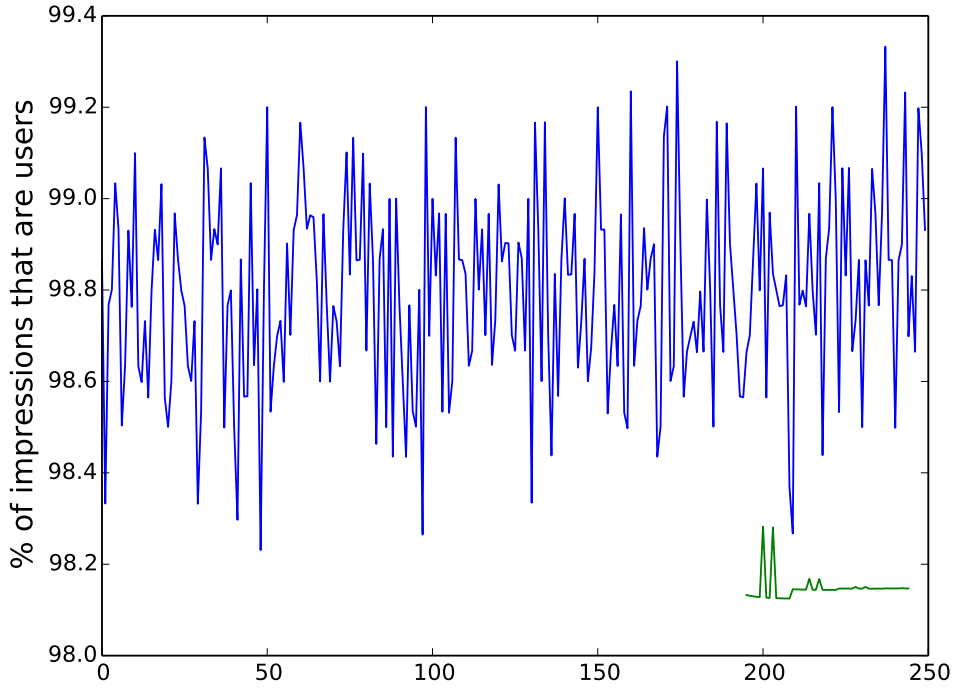

Figure C.53: Arima Allow Drift True - Uniques Percentage forecast from 2013-12-31 12:00:00 to 2014-01-25 00:00:00

σ (Real Data)	RMSE	MASE
0.22	0.72	0.7065

Table C.53: Arima Allow Drift True - Error for Uniques Percentage forecast from 2013-12-31 12:00:00 to 2014-01-25 00:00:00

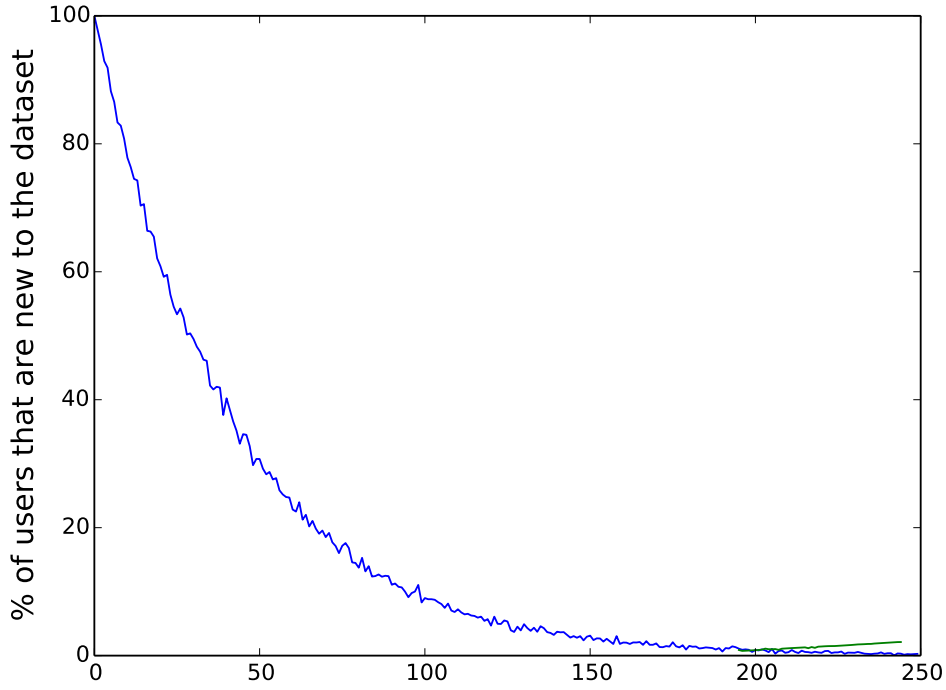

Figure C.54: Arima Allow Drift True - New uniques forecast from 2013-12-31 12:00:00 to 2014-01-25 00:00:00

σ (Real Data)	RMSE	MASE
0.25	1.05	0.2602

Table C.54: Arima Allow Drift True - Error for New Uniques forecast from 2013-12-31 12:00:00 to 2014-01-25 00:00:00

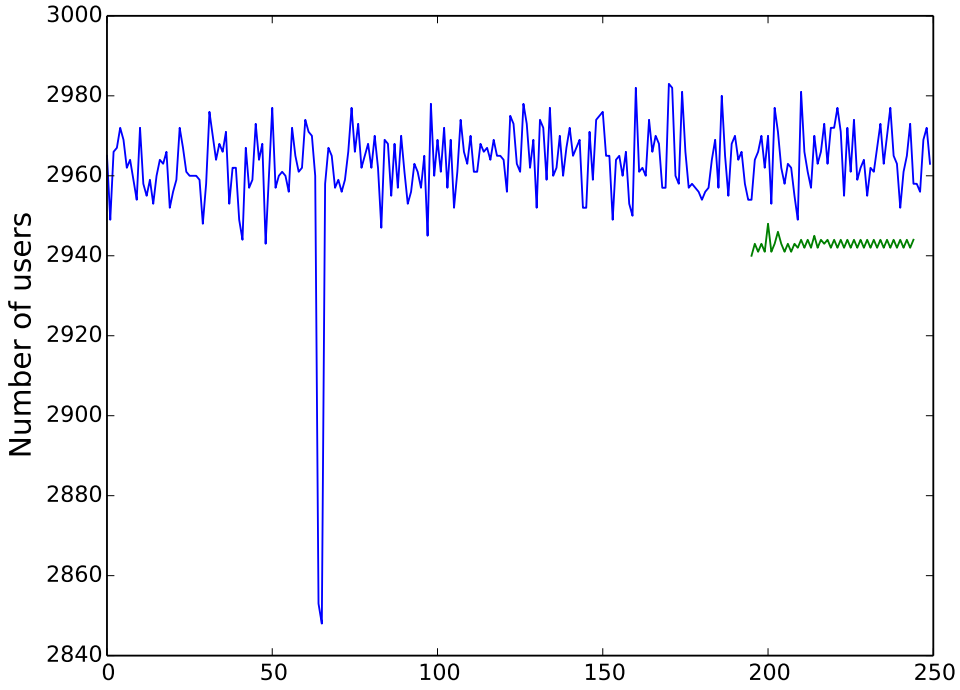

Figure C.55: Arima Allow Drift True - Uniques calculated using percentages forecast from 2013-12-31 12:00:00 to 2014-01-25 00:00:00

σ (Real Data)	RMSE	MASE
7.13	22.92	0.5941

Table C.55: Arima Allow Drift True - Error for Uniques calculated using percentages forecast from 2013-12-31 12:00:00 to 2014-01-25 00:00:00

C.12 Arima Allow Drift False - 12h

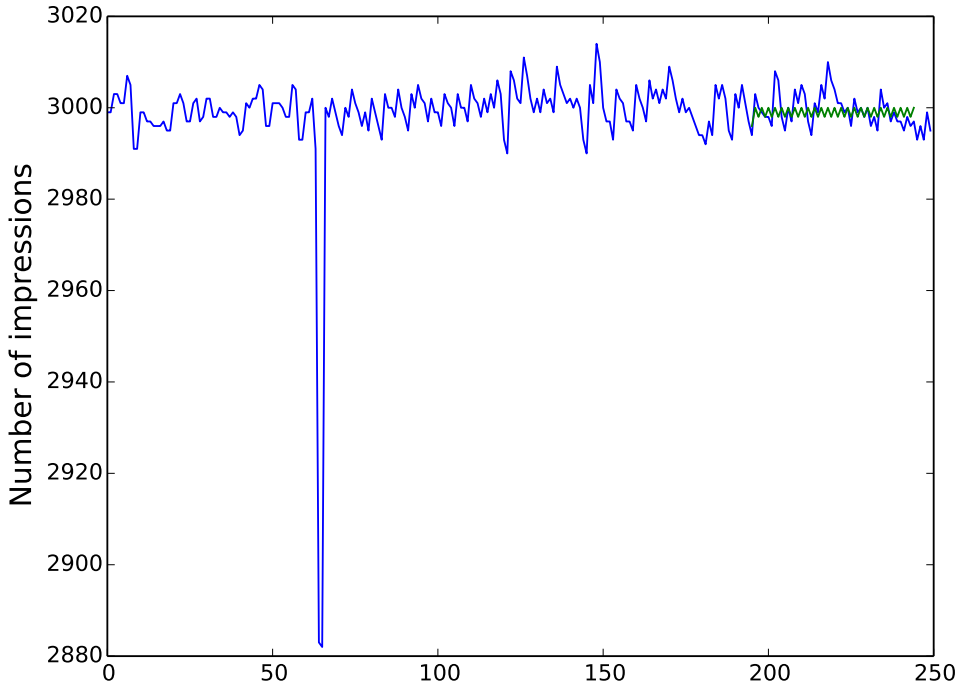

Figure C.56: Arima Allow Drift False - Impressions forecast from 2013-12-31 12:00:00 to 2014-01-25 00:00:00

σ (Real Data)	RMSE	MASE
3.76	3.43	0.1439

Table C.56: Arima Allow Drift False - Error for Impressions forecast from 2013-12-31 12:00:00 to 2014-01-25 00:00:00

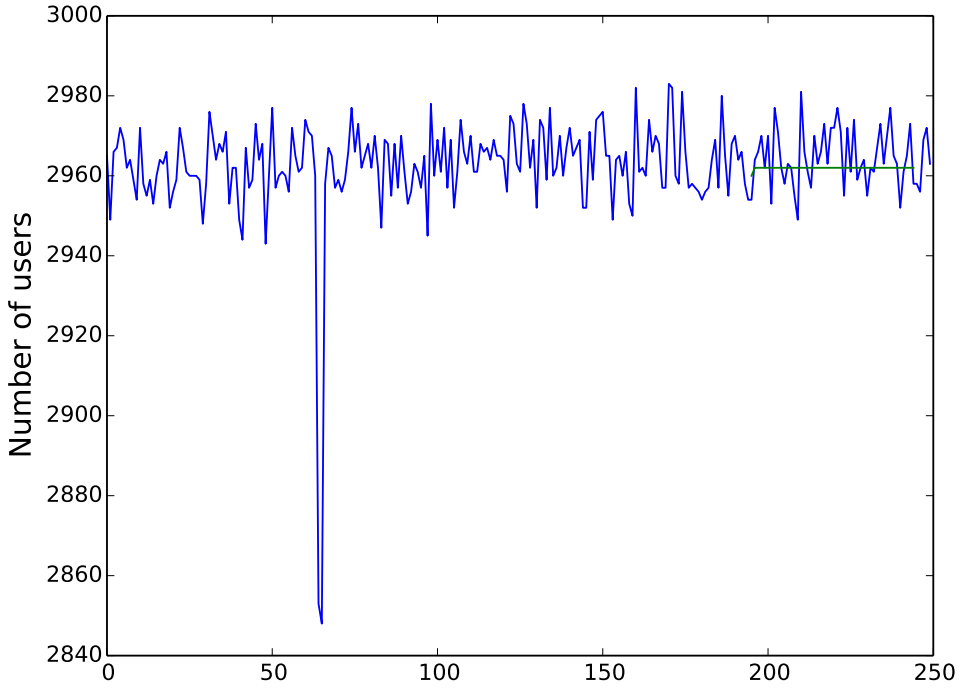

Figure C.57: Arima Allow Drift False - Uniques forecast from 2013-12-31 12:00:00 to 2014-01-25 00:00:00

σ (Real Data)	RMSE	MASE
7.13	7.7	0.1631

Table C.57: Arima Allow Drift False - Error for Uniques forecast from 2013-12-31 12:00:00 to 2014-01-25 00:00:00

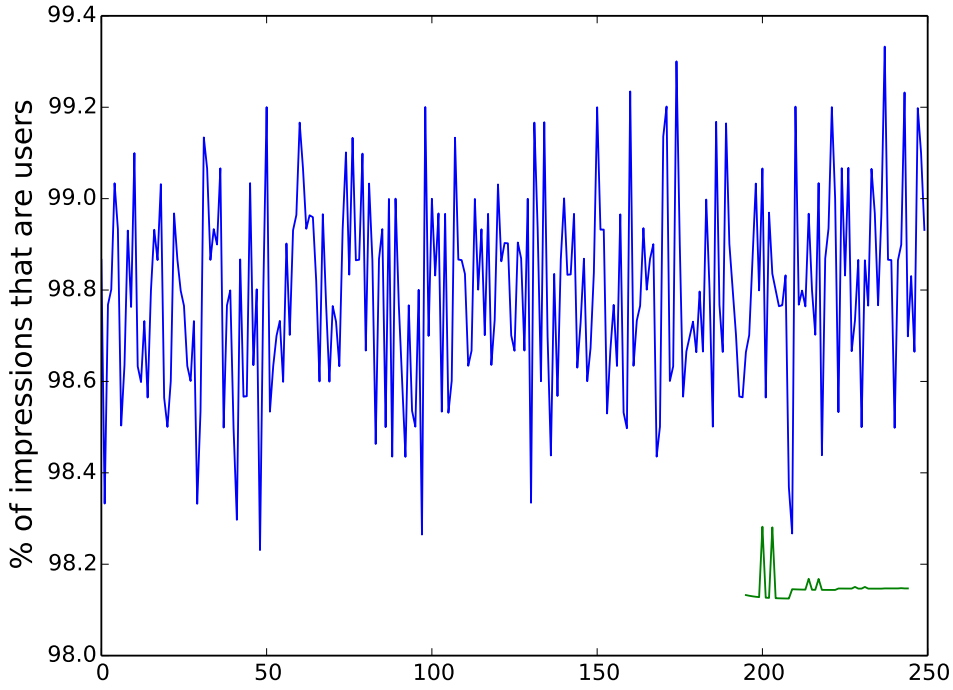

Figure C.58: Arima Allow Drift False - Uniques Percentage forecast from 2013-12-31 12:00:00 to 2014-01-25 00:00:00

σ (Real Data)	RMSE	MASE
0.22	0.72	0.7065

Table C.58: Arima Allow Drift False - Error for Uniques Percentage forecast from 2013-12-31 12:00:00 to 2014-01-25 00:00:00

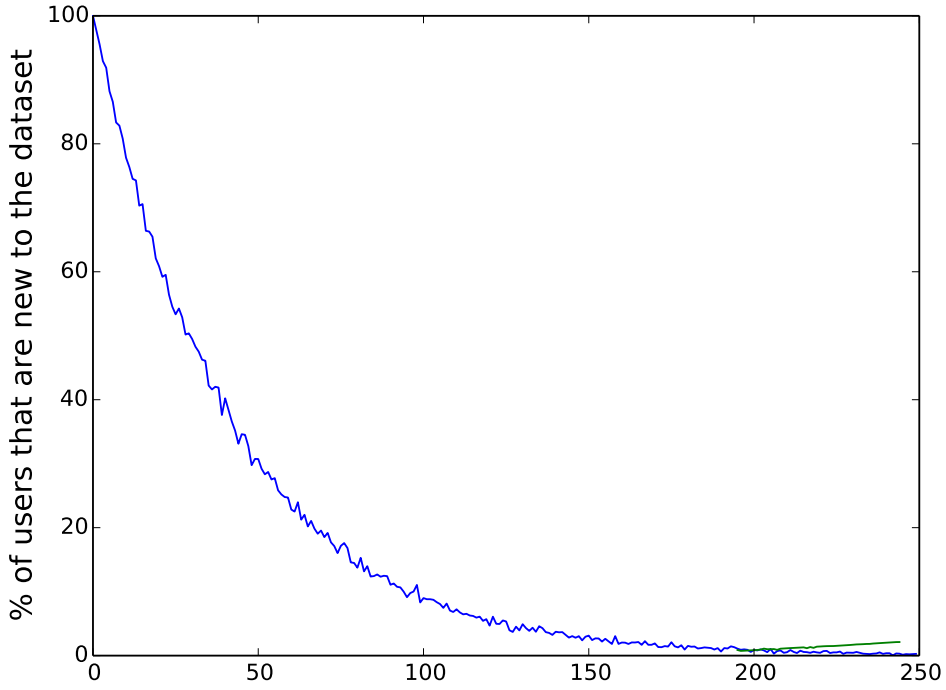

Figure C.59: Arima Allow Drift False - New uniques forecast from 2013-12-31 12:00:00 to 2014-01-25 00:00:00

σ (Real Data)	RMSE	MASE
0.25	1.05	0.2602

Table C.59: Arima Allow Drift False - Error for New Uniques forecast from 2013-12-31 12:00:00 to 2014-01-25 00:00:00

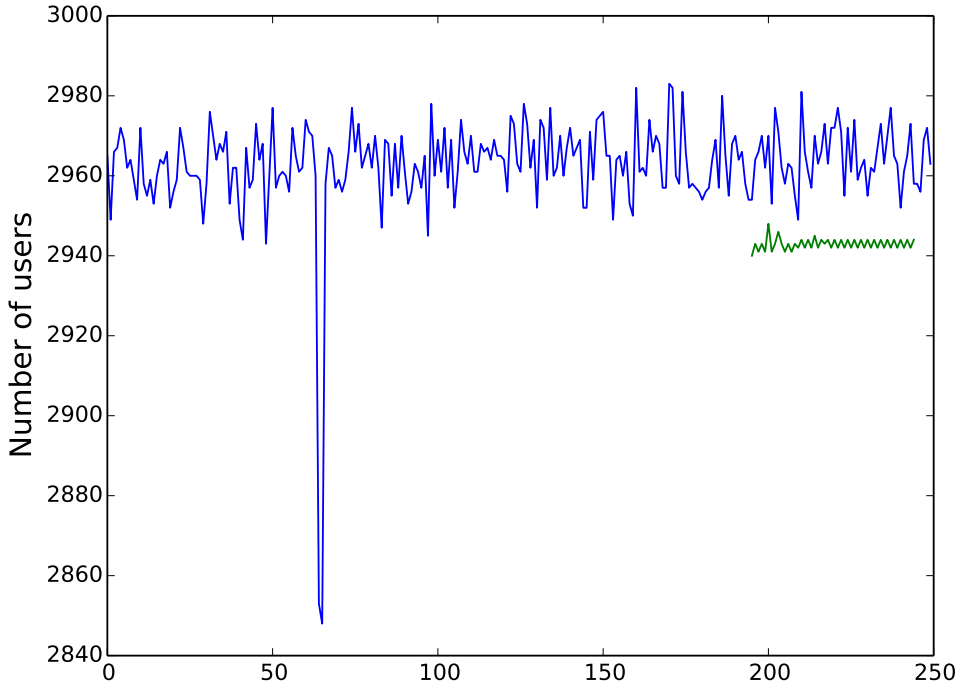

Figure C.60: Arima Allow Drift False - Uniques calculated using percentages forecast from 2013-12-31 12:00:00 to 2014-01-25 00:00:00

σ (Real Data)	RMSE	MASE
7.13	22.92	0.5941

Table C.60: Arima Allow Drift False - Error for Uniques calculated using percentages forecast from 2013-12-31 12:00:00 to 2014-01-25 00:00:00

C.13 Baseline - 24h

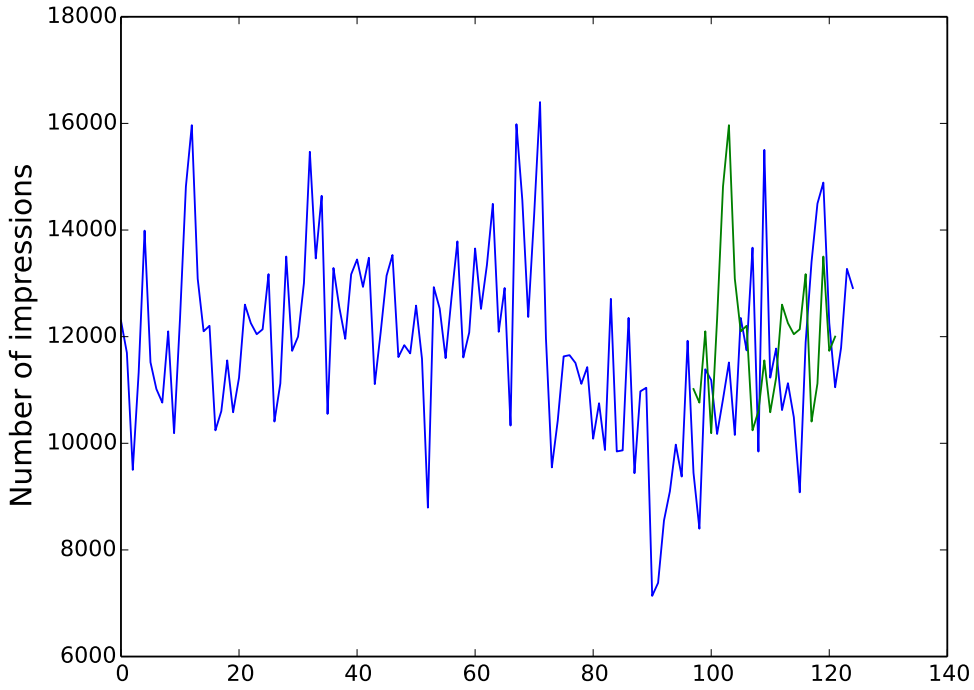

Figure C.61: Baseline - Impressions forecast from 2013-12-31 00:00:00 to 2014-01-24 00:00:00

σ (Real Data)	RMSE	MASE
1686.92	2275.91	0.3304

Table C.61: Baseline - Error for Impressions forecast from 2013-12-31 00:00:00 to 2014-01-24 00:00:00

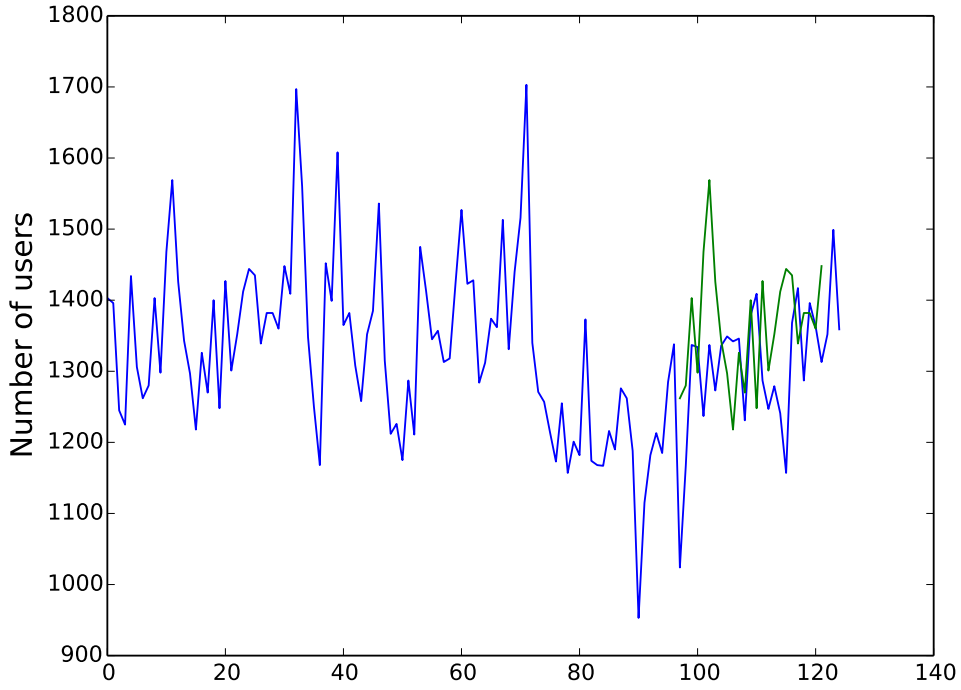

Figure C.62: Baseline - Uniques forecast from 2013-12-31 00:00:00 to 2014-01-24 00:00:00

σ (Real Data)	RMSE	MASE
91.68	131.08	0.2772

Table C.62: Baseline - Error for Uniques forecast from 2013-12-31 00:00:00 to 2014-01-24 00:00:00

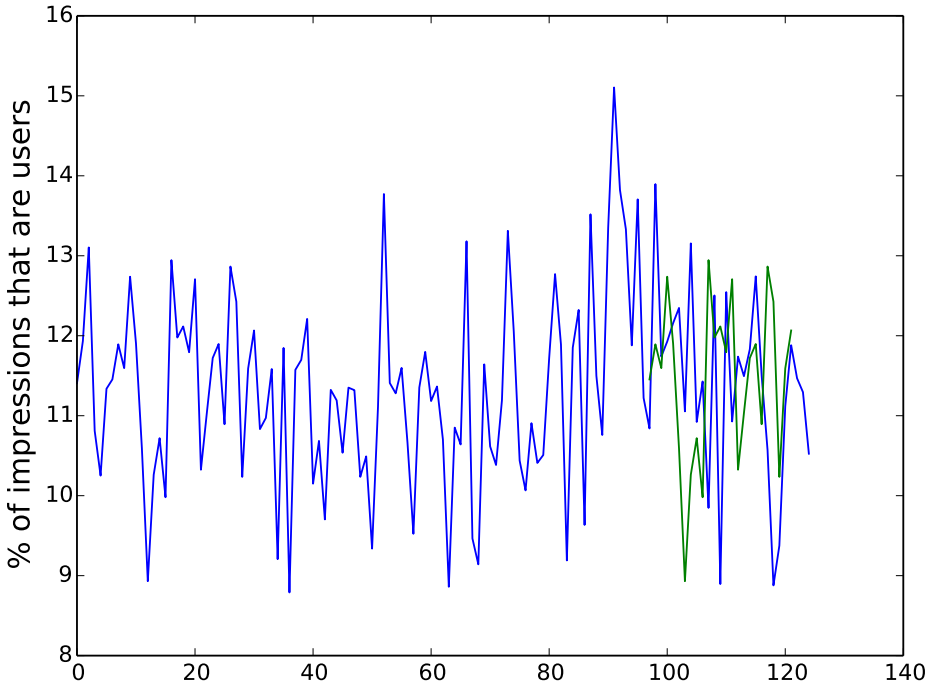

Figure C.63: Baseline - Uniques Percentage forecast from 2013-12-31 00:00:00 to 2014-01-24 00:00:00

σ (Real Data)	RMSE	MASE
1.16	1.67	0.2623

Table C.63: Baseline - Error for Uniques Percentage forecast from 2013-12-31 00:00:00 to 2014-01-24 00:00:00

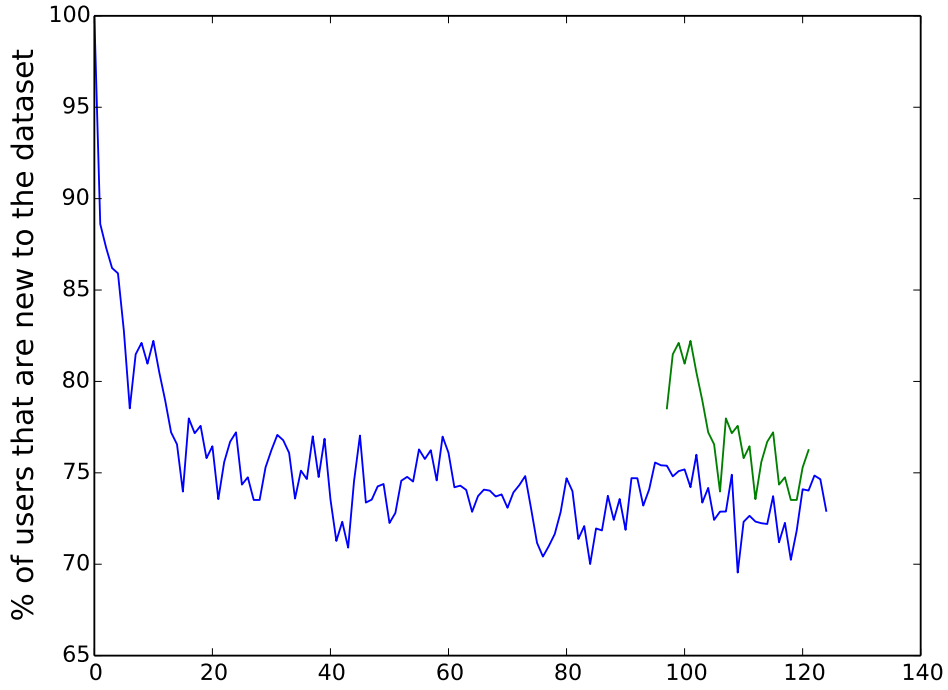

Figure C.64: Baseline - New uniques forecast from 2013-12-31 00:00:00 to 2014-01-24 00:00:00

σ (Real Data)	RMSE	MASE
1.55	4.4	0.6787

Table C.64: Baseline - Error for New Uniques forecast from 2013-12-31 00:00:00 to 2014-01-24 00:00:00

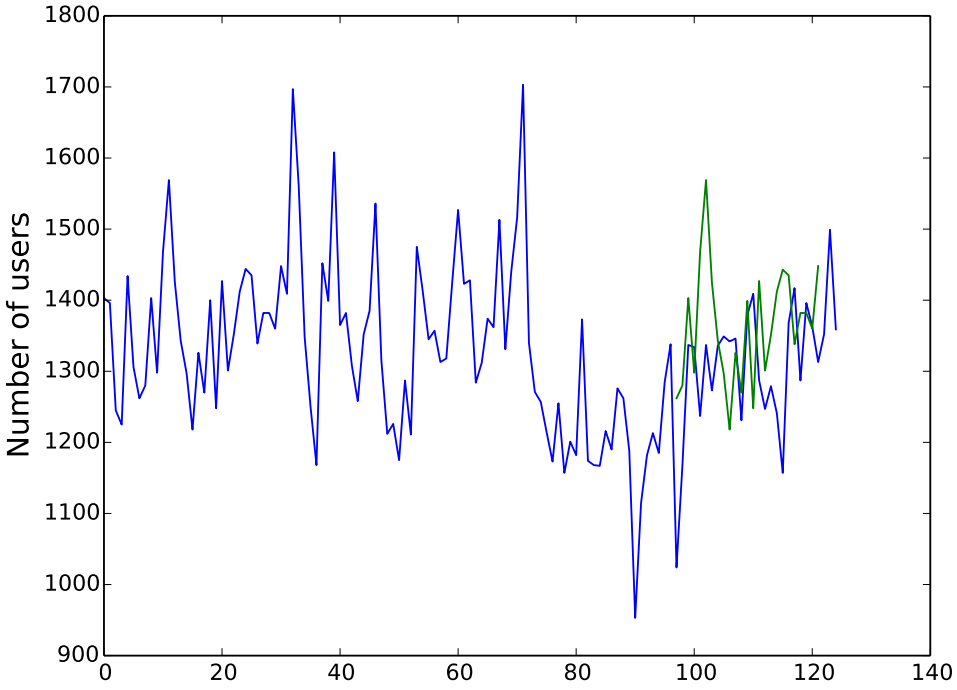

Figure C.65: Baseline - Uniques calculated using percentages forecast from 2013-12-31 00:00:00 to 2014-01-24 00:00:00

σ (Real Data)	RMSE	MASE
91.68	130.94	0.2769

Table C.65: Baseline - Error for Uniques calculated using percentages forecast from 2013-12-31 00:00:00 to 2014-01-24 00:00:00

C.14 Arima Allow Drift True - 24h

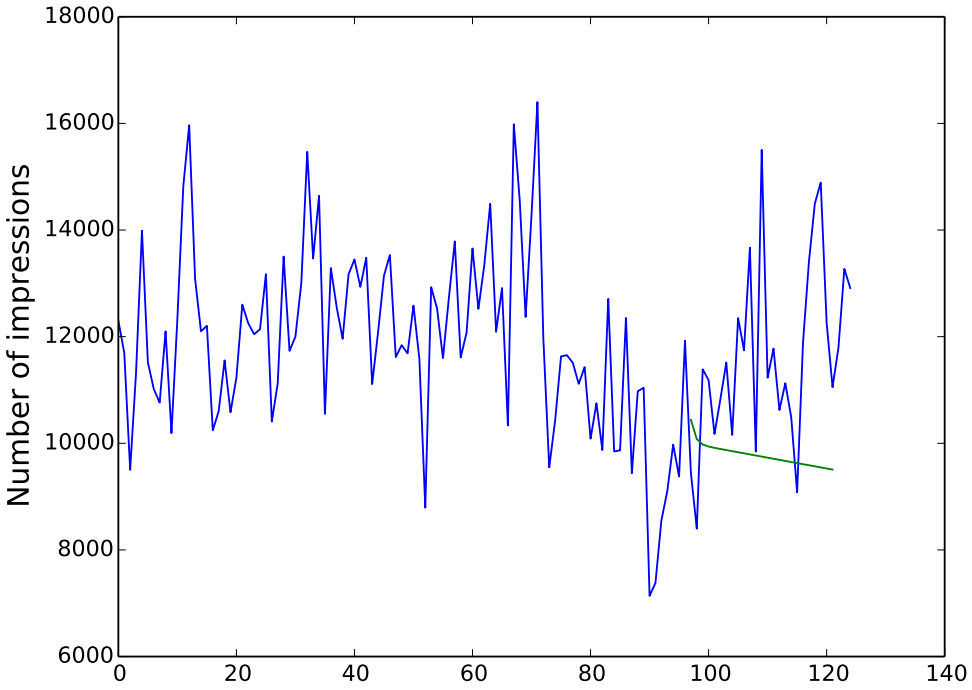

Figure C.66: Arima Allow Drift True - Impressions forecast from 2013-12-31 00:00:00 to 2014-01-24 00:00:00

σ (Real Data)	RMSE	MASE
1686.92	2547.73	0.3525

Table C.66: Arima Allow Drift True - Error for Impressions forecast from 2013-12-31 00:00:00 to 2014-01-24 00:00:00

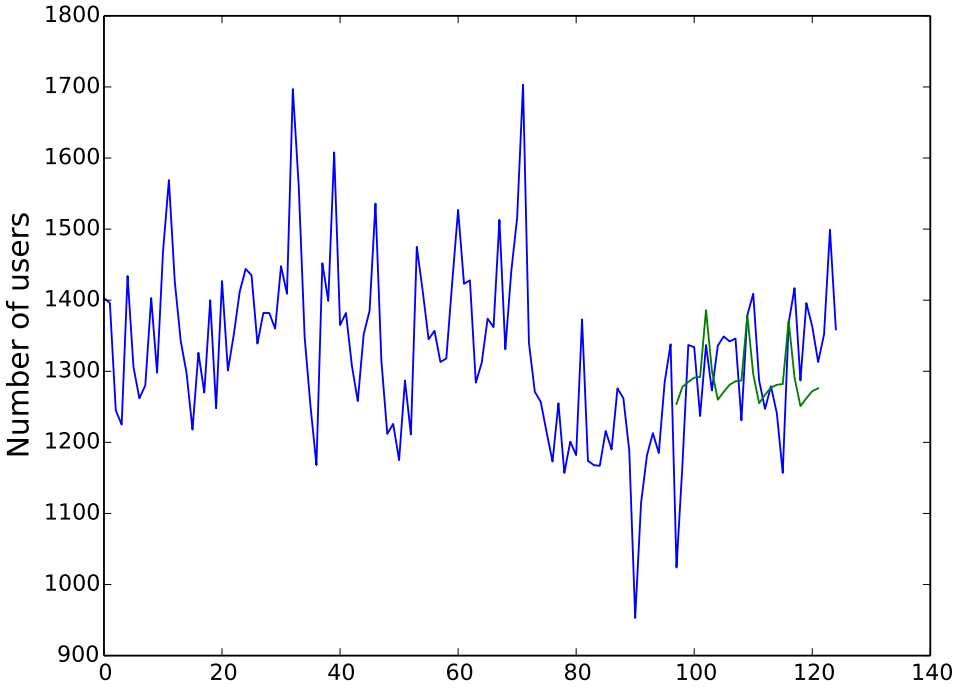

Figure C.67: Arima Allow Drift True - Uniques forecast from 2013-12-31 00:00:00 to 2014-01-24 00:00:00

σ (Real Data)	RMSE	MASE
91.68	83.55	0.1761

Table C.67: Arima Allow Drift True - Error for Uniques forecast from 2013-12-31 00:00:00 to 2014-01-24 00:00:00

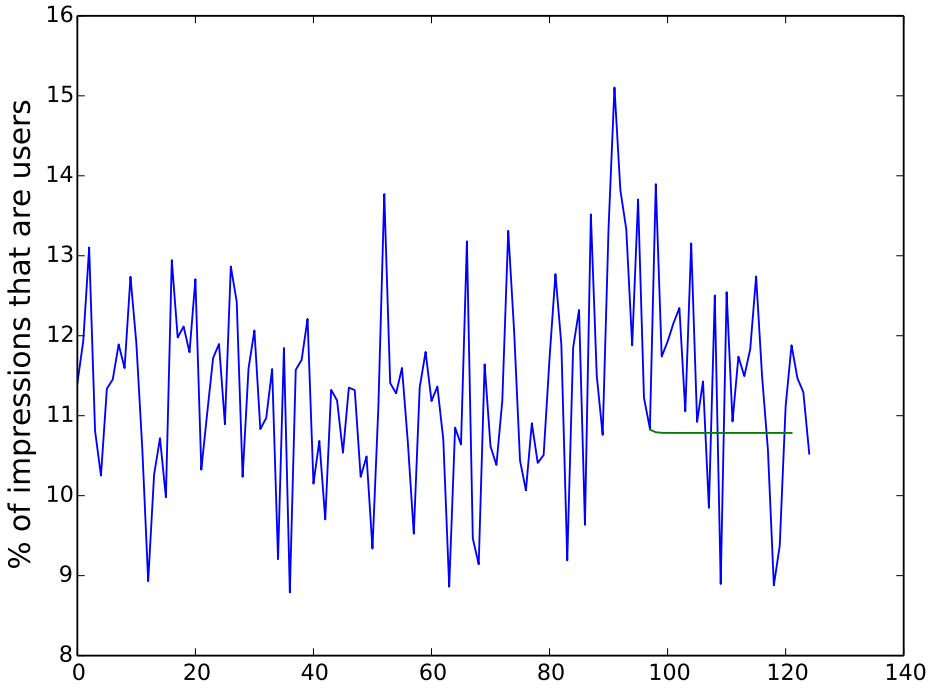

Figure C.68: Arima Allow Drift True - Uniques Percentage forecast from 2013-12-31 00:00:00 to 2014-01-24 00:00:00

σ (Real Data)	RMSE	MASE
1.16	1.37	0.2294

Table C.68: Arima Allow Drift True - Error for Uniques Percentage forecast from 2013-12-31 00:00:00 to 2014-01-24 00:00:00

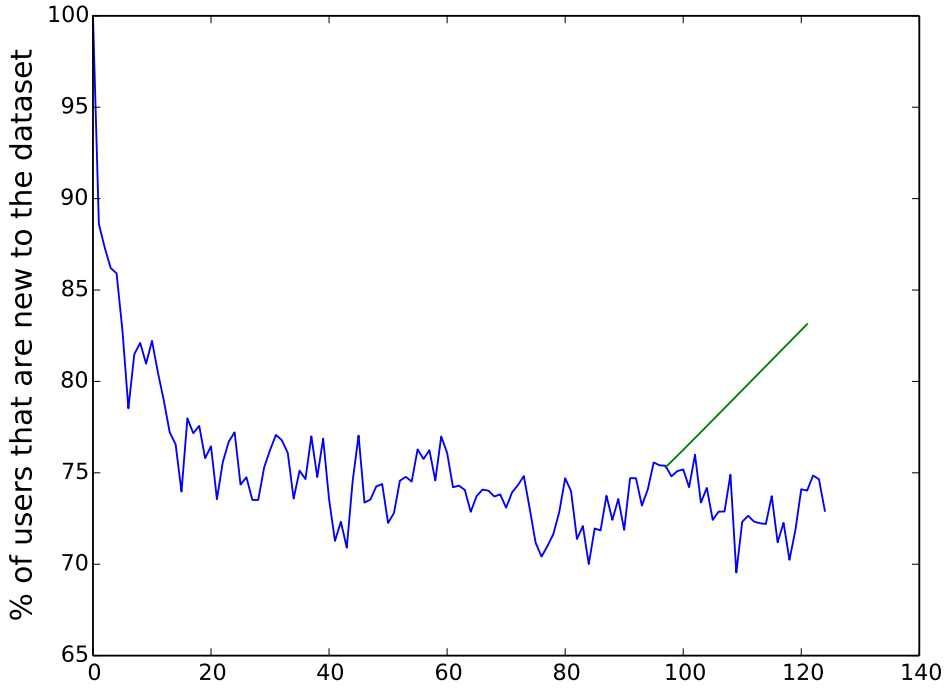

Figure C.69: Arima Allow Drift True - New uniques forecast from 2013-12-31 00:00:00 to 2014-01-24 00:00:00

σ (Real Data)	RMSE	MASE
1.55	6.95	1.0377

Table C.69: Arima Allow Drift True - Error for New Uniques forecast from 2013-12-31 00:00:00 to 2014-01-24 00:00:00

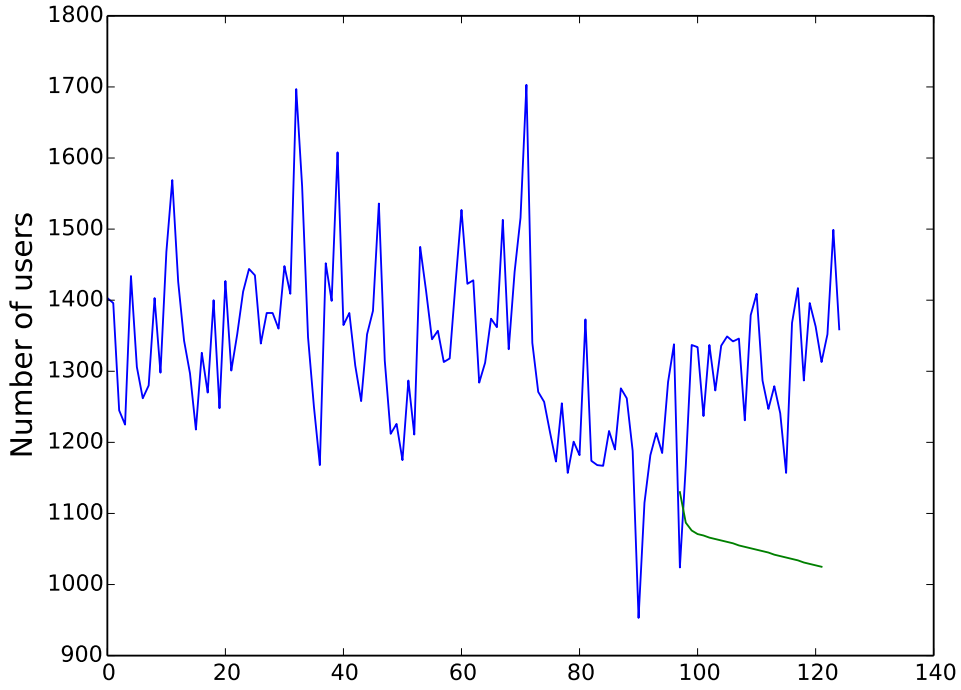

Figure C.70: Arima Allow Drift True - Uniques calculated using percentages forecast from 2013-12-31 00:00:00 to 2014-01-24 00:00:00

σ (Real Data)	RMSE	MASE
91.68	264.97	0.6715

Table C.70: Arima Allow Drift True - Error for Uniques calculated using percentages forecast from 2013-12-31 00:00:00 to 2014-01-24 00:00:00

C.15 Arima Allow Drift False - 24h

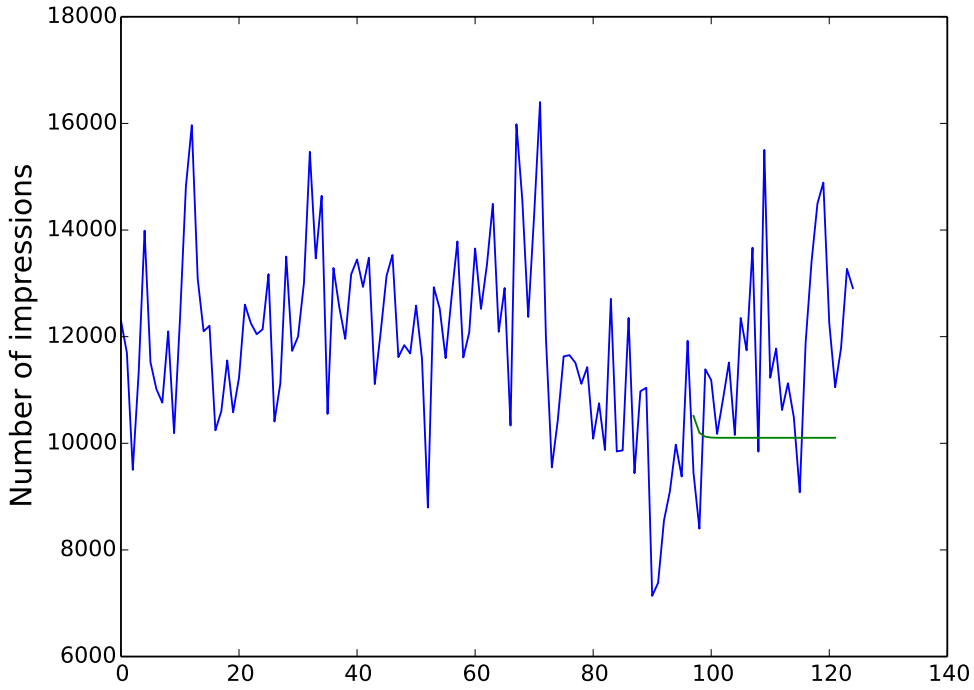

Figure C.71: Arima Allow Drift False - Impressions forecast from 2013-12-31 00:00:00 to 2014-01-24 00:00:00

σ (Real Data)	RMSE	MASE
1686.92	2260.99	0.3044

Table C.71: Arima Allow Drift False - Error for Impressions forecast from 2013-12-31 00:00:00 to 2014-01-24 00:00:00

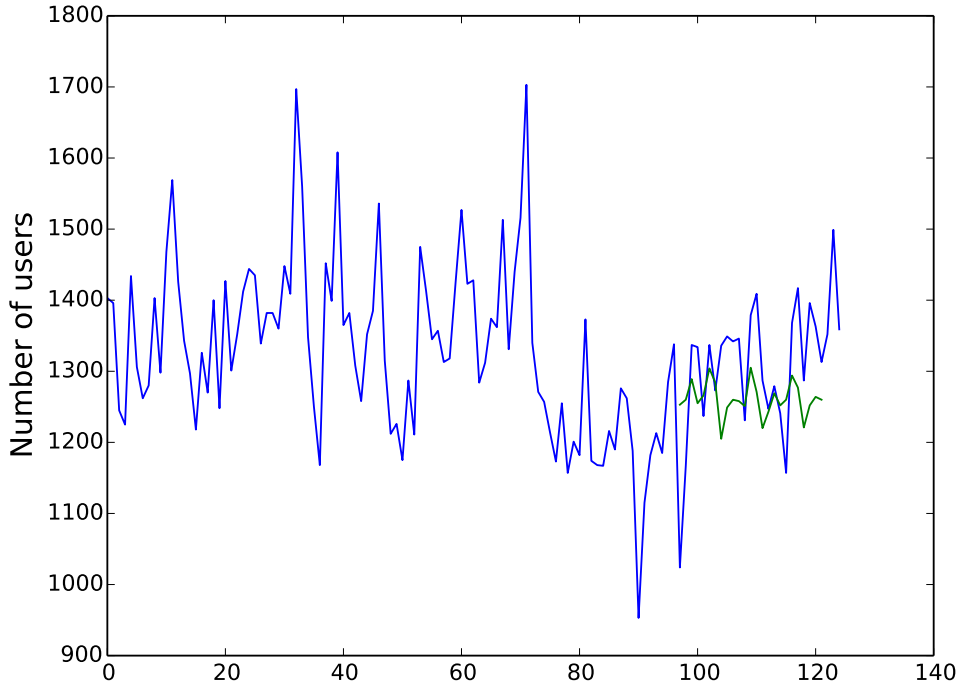

Figure C.72: Arima Allow Drift False - Uniques forecast from 2013-12-31 00:00:00 to 2014-01-24 00:00:00

σ (Real Data)	RMSE	MASE
91.68	93.03	0.2046

Table C.72: Arima Allow Drift False - Error for Uniques forecast from 2013-12-31 00:00:00 to 2014-01-24 00:00:00

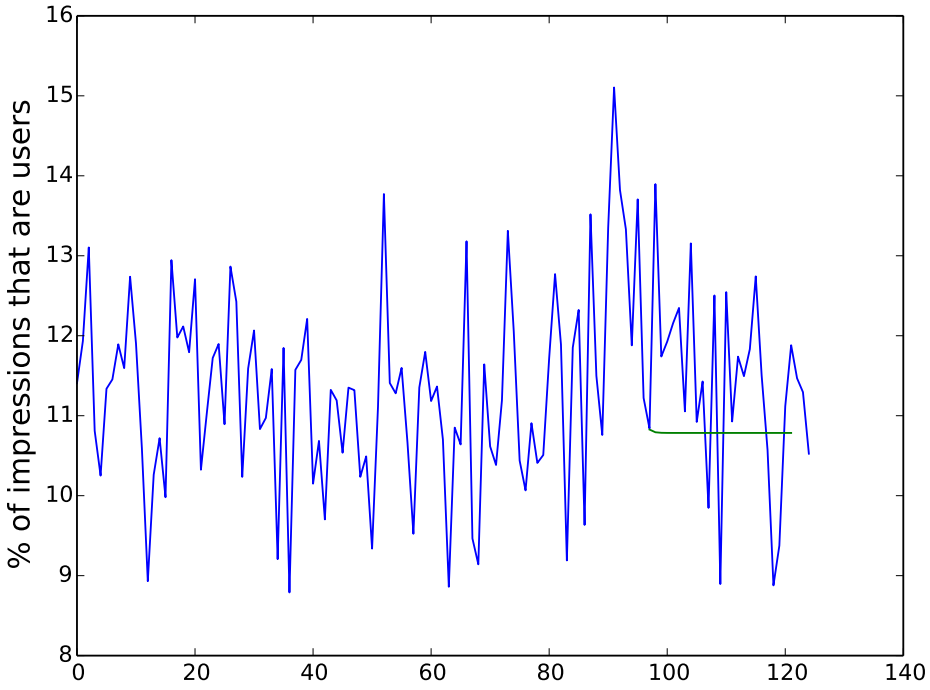

Figure C.73: Arima Allow Drift False - Uniques Percentage forecast from 2013-12-31 00:00:00 to 2014-01-24 00:00:00

σ (Real Data)	RMSE	MASE
1.16	1.37	0.2294

Table C.73: Arima Allow Drift False - Error for Uniques Percentage forecast from 2013-12-31 00:00:00 to 2014-01-24 00:00:00

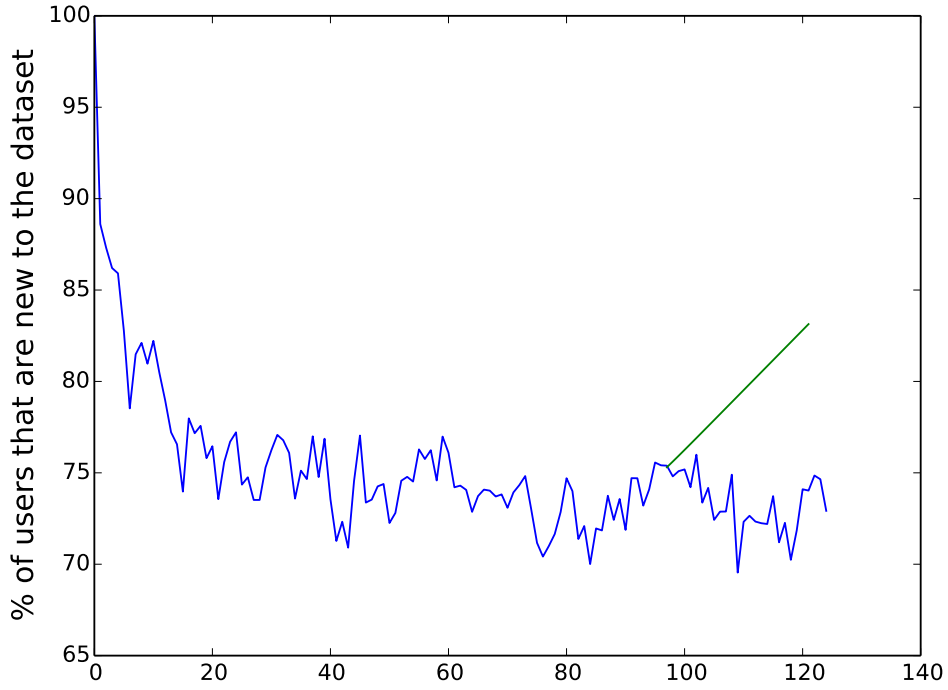

Figure C.74: Arima Allow Drift False - New uniques forecast from 2013-12-31 00:00:00 to 2014-01-24 00:00:00

σ (Real Data)	RMSE	MASE
1.55	6.95	1.0377

Table C.74: Arima Allow Drift False - Error for New Uniques forecast from 2013-12-31 00:00:00 to 2014-01-24 00:00:00

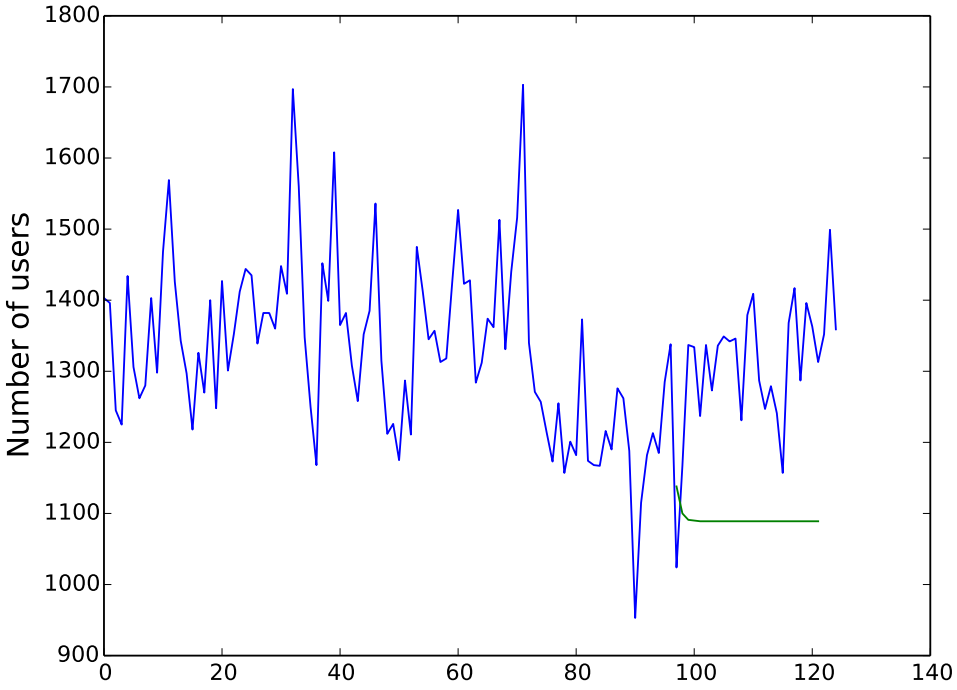

Figure C.75: Arima Allow Drift False - Uniques calculated using percentages forecast from 2013-12-31 00:00:00 to 2014-01-24 00:00:00

σ (Real Data)	RMSE	MASE
91.68	227.36	0.5731

Table C.75: Arima Allow Drift False - Error for Uniques calculated using percentages forecast from 2013-12-31 00:00:00 to 2014-01-24 00:00:00